FOPPA: A database of French Open Public Procurement Award notices

Lucas Potin, Vincent Labatut, Rosa Figueiredo, Christine Largeron, Pierre-Henri Morand

AVIGNON UNIVERSITÉ

FOPPA

A database of French Open Public Procurement Award notices

Lucas Potin[1,4]
Vincent Labatut[1,4]
Rosa Figueiredo[1,4]
Christine Largeron[2]
Pierre-Henri Morand[3,4]

02/10/2022
Revised 26/03/2024

[1] Laboratoire Informatique d'Avignon – LIA UPR 4128
{firstname.lastname}@univ-avignon.fr

[2] Laboratoire Hubert Curien – LHC UMR 5516
christine.largeron@univ-st-etienne.fr

[3] Laboratoire Biens, Normes, et Contrats – LBNC UPR 3788
pierre-henri.morand@univ-avignon.fr

[4] Agorantic – FR 3621

anr©

LABORATOIRE INFORMATIQUE D'AVIGNON

LBNC

Contents

This document presents the FOPPA database, as well as the process applied to initialize it based on European public open data.

Context This work takes place in the context of the DeCoMaP ANR Project[1] (ANR-19-CE38-0004), which aims at automating the detection of fraud in public procurement.

Resources The source code v1.0.3 that implements the process described in this report is publicly available online, on GitHub, at the following URL:

https://github.com/CompNet/FoppaInit/tree/v1.0.3 .

The database v1.1.3 resulting from this process is also publicly available, as a Zenodo repository, at the following URL:

https://doi.org/10.5281/zenodo.10879932

Citation If you use this database, please cite the associated data paper:

- L. Potin, V. Labatut, C. Largeron, and P. H. Morand. "FOPPA: an open database of French public procurement award notices from 2010–2020". In: *Scientific Data* 10 (2023), p. 303. DOI: 10.1038/s41597-023-02213-z

and possibly the present report:

- L. Potin, V. Labatut, R. Figueiredo, C. Largeron, and P.-H. Morand. *FOPPA: a database of French Open Public Procurement Award notices*. Tech. rep. Avignon Université, 2022. URL: https://hal.archives-ouvertes.fr/hal-03796734

Acknowledgments Most of the source code was written by Lucas Potin in the context of his PhD work. In addition, several students from the CERI (*Centre d'Enseignement et de Recherche en Informatique* – Center for Computer Science of Avignon University) punctually participated on certain specific points. During their 5th year *Business Intelligence* class, Yanis Labrak, Quentin Raymondaud, and Philippe Turcotte helped to connect the FOPPA database to the BRÉF (see Section 6.3.2). During her 3rd year internship, Rim Amarat worked on improving the separation of multiple criteria and the categorization of criteria.

Formatting We adopt the following conventions in this report. values extracted from a database and table names are written using a `monospaced font`.

Examples extracted from the raw data are represented using a specific background color:

```
<Raw Data>
```

Organization We first summarize the main notions related to public procurement in Section 1. We then turn to our main data source, the TED, which we describe in Section 2, focusing on the problems that we identified in this database.

In the rest of the document, we describe the methods that we propose to solve these problems. We start with two minor issues in Section 3, regarding the separation of agents and criteria. Then, in Section 4, we deal with the major problem, which is about agent identification. Finally, in Section 5, we perform a post-processing aiming at improving the quality of our data.

We assess the effect of each one of these steps upon the database. We conclude in Section 6 by summarizing the characteristics of the FOPPA database, identifying the issues that remain to be solved, and discussing the next steps of our work in the context of DeCoMaP.

[1] https://anr.fr/Projet-ANR-19-CE38-0004

1 Public Procurement and Related Notions

Public procurement refers to the purchase of goods, services and works by a public authority (the buyer) from a legal entity governed by public or private law (the supplier). In this work, we focus on the case of French public procurement.

In French law, a *public authority* is a public or private buyer, which belongs to one of three possible categories[2]:

- Legal persons governed by public law;
- Legal persons governed by private law, pursuing a mission of public interest and controlled or funded predominantly by public funds;
- Legal persons governed by private law, constituted of public authorities, and aiming at conducting certain collective activities.

This includes, but is not limited to, public governments and state-owned enterprises. The acronym CAE (Contract Authority or Entity) is used to refer to buyers.

A *public entity* is a specific type of public authority acting as a network operator, i.e. operating in certain particular activity domains related to water or energy networks.

Public procurement must follow a specific set of rules defined by law, and aiming at respecting three principles:

- *Freedom of access*: all potential candidates must be able to access the necessary information;
- *Level playing field*: all candidates must be treated equally by the public authority;
- *Transparency*: the awarding procedure and its outcome must be provided to all candidates.

1.1 Awarding Process

The general steps of the process consisting in awarding a contract to a supplier are as follows[3]:

1. Identification of the buyer's needs;
2. Breakdown of these needs in several parts;
3. Estimation of the value of each lot;
4. Selection of the most appropriate procedure (see below);
5. Precise specification of the needs taking the form of a public contract;
6. Advertisement for the public contract;
7. Selection of the best offer, which is awarded to a supplier;
8. Entering into a contract and conclusion of the process.

Lots & Activity Domain At Step 2, the public authority may separate its needs into several parts called *lots*. Each lot is associated to one or several codes expressed using the CPV system (Common Procurement Vocabulary[4]) defined by the European Union. Each such code describes the main or secondary subjects of a contract, e.g. *Fruit seeds*, *Insulation work*.

Procedure & Thresholds The *procedure* that must be followed at Step 4 is an important aspect of public procurement. It depends mainly on the value estimated at Step 3, but also on other factors [2]: nature and activity domain of the public authority (state, local government, health institution...), and nature of the contract (goods, services, works).

Under certain very specific conditions, it is possible to use a negotiated procedure without a prior call for competition (*Procédure négociée sans mise en concurrence*) [2], noted

[2] https://www.economie.gouv.fr/daj/pouvoirs-adjudicateurs-et-entites-adjudicatrices-2019

[3] https://organisme-de-formation-professionnelle.fr/2019/04/08/marche-public-appel-d-offres-definition-deroulement/

[4] https://simap.ted.europa.eu/web/simap/cpv

NOC/NOP. These conditions include the occurrence of an emergency situation, the absence of any reasonably acceptable offer, and the case where the needs are so specific that they can be fulfilled only by a single supplier.

But in the regular case, the factors mentioned above are used to determine a so-called *European Threshold*, that ranges from 139 k€ to 5.35 M€ for the 2020–2021 period. These thresholds are revised every two years by the European Commission. We detail them in Table 32, in the appendix. If the estimated value of the contract is *below* this European threshold, the public authority must follow the so-called *Adapted Procedure* (Section 1.2). If it is *above* this threshold, it must follow the more constraining *Formalized Procedure* (Section 1.3).

Advertisement & Notices The way the contract is advertised at Step 6 completely depends on the selected procedure. Most of the time, it is a call for tenders, published as a *contract notice*, that ends after a so-called *acceptance period*. Once the various offers have been studied at Step 7, the public authority decides whether or not to award the different lots to one or more candidates, who are called *winners*. The public authority indicates its choice with a *contract award notice*, which is the formal document providing the details regarding the contract attribution. The criteria used to select the winner are an important part of the process, which we discuss in Section 1.4.

Framework Agreements These are a specific type of contract between one or several buyers and suppliers, that aim at establishing some general rules regarding their commercial transactions for a specific period, in particular in terms of prices and quantities. There are two types of framework agreement, depending on how these future transactions are handled. If they are considered as *subsequent markets*, then it is possible for the buyer to put the suppliers back into competition, which requires a call for tenders. The transactions can also take the form of *purchase orders*, which do not require any call for tender, but must apply the constraints contractually defined in the framework agreement.

Unlike the framework agreements themselves, the contracts and purchase orders conducted *within* a framework agreement do not need to be advertised. Put differently, they are not necessarily described through a contract notice. Similarly, the law does not require advertising the result of the awarding process, i.e. to publish a contract award notice (it is possible to do so, though).

Correction & Cancelation It is possible for the public authority to *correct* a notice, before the acceptance period of the offers is over[5]. Such a correction can aim at fixing some errors in the original notice, but also at changing the conditions for awarding. If these changes are significant, they must be published as a specific *correction notice*, using the same outlet as the initial notice. Such a correction may result in the extension of the acceptance period.

Finally, it is also possible for the public authority to cancel a contract. There are two main reasons for this[6]: *unsuccessful procedure*, and *public interest*. The former covers the cases where no tender is made, or all of them are inappropriate (in particular: not meeting all requirements of the call, or failing to provide all the required information). The latter covers economic (e.g. the CAE has exhausted its budget), judicial (e.g. the CAE detected some irregularity in the procedure) and technical (e.g. the CAE identified a mistake in the technical requirement listed in the call) motives. The public authority can advertise the cancellation (typically, on the TED), but this is optional, not compulsory. However, it must inform the bidders involved in the process. Also, if it decides to re-tender the market, it must indicate this in the new contract notice.

[5] http://www.marchespublicspme.com/avant-la-reponse/lexique-des-termes-de-marches-publics/actualites/2020/12/29/avis-rectificatif-dans-les-marches-publics-qu-est-ce-que-c-est_15704.html

[6] https://www.economie.gouv.fr/daj/abandon-procedure-2019

1.2 Adapted Procedure

The adapted procedure, or `MAPA` (*Marché A Procédure Adaptée* – Adapted procedure market) leaves it up to the public authority to choose the conditions of attribution of the contract, provided the three principles mentioned before are respected (freedom of access, level playing field, transparency). However, additional thresholds control the way the contract must be advertised.

Below 40 k€ It is not compulsory to publicly advertise the procurement or to perform any competitive call.

Above 40 k€ It is compulsory to publicly advertise the contract. The advertisement medium depends on the estimated value of the contract[7]:

- **Between 40 k€ and 90 k€**: the contract must be advertised by whatever means the buyer wants to use.
- **Between 90 k€ and the European Threshold**: the contract must be advertised in the BOAMP (*Bulletin Officiel des Annonces de Marchés Publics* – French official bulletin of public procurement notices).

For the sake of completeness, let us mention that social and special services have a *specific status* that allows them to use different thresholds [2].

1.3 Formalized Procedure

Above the European Threshold, the public authority must advertise the contract through the BOAMP and the OJEU (Official Journal of the European Union). The online publication outlet of the OJEU that is dedicated to the publication of public procurement notices is called the TED[8] (*Tenders Electronic Daily*), and we discuss it later in Section 2.1.

The public authority can choose between four types of formalized procedures, and must stick to the selected procedure until a winner has been identified [2].

Open Procedure (noted `OPE`) The public authority publishes a call for tenders. Any interested candidate can submit a bid. This procedure is generally used when the needs are straightforward, the award process is simple, and the public authority expects only a few number of candidates.

Restricted Procedure (noted `RES`) The public authority also publishes a call for tenders, but only the candidates pre-selected by the public authority can submit a bid. It is two-stepped: first, the potential candidates are asked to express an interest to the contract under the form of a preliminary file; second, the public authority establishes a short list of candidates that are allowed to submit a full bid. This procedure is used for complex contracts and/or when many candidates are expected.

Competitive Dialogue (noted `COD`) The first step of the restricted procedure is applied iteratively, each candidate being able to revise its bid. The public authority can discard some candidates at each iteration. When the public authority is satisfied, it invites the remaining candidates to submit a full bid. This procedure is used for complex contracts, in particular when the needs cannot be identified clearly in advance.

Competitive Procedure with Negotiation (noted `NIC/NIP`) This procedure is similar to competitive dialogue, except the public authority can decide not to negotiate, depending on the nature of the preliminary bids.

[7] https://www.service-public.fr/professionnels-entreprises/vosdroits/F23371
[8] https://ted.europa.eu/

1.4 Award Criteria

The public authority has to specify in advance which criteria will be used to select the winning bid. They must respect the following principles[9]:

- Allow selecting the most economically advantageous tender;
- Only apply to the bid, not the candidate itself;
- Be fair and sufficiently precise;
- Be specified before the call for tenders;
- Must be either weighted or prioritized.

The law does not explicitly list all possible criteria, but rather proposes several categories of criteria, and sets some boundaries. Some criteria defined at the national level and used with the adapted procedure are illicit at the European level and cannot be used with the formalized procedure.

In case of formalized procedure (cf. Section 1.3, each criterion must be associated to a weight, that allows assessing its importance relative to the other criteria.

It is possible to use a *single* criterion, in which case it is necessarily the *contract value*. However, this is allowed only if the contract aims at buying goods or services that are standardized, and whose quality can vary from one supplier to the other.

Otherwise, the public authority has to use several criteria, which must be related to the object of the contract or its implementation, and must include the contract value. The other possible criteria are organized in the following categories.

Quality The notion of quality covers various aspects of the bid: technical value, aesthetic and functional characteristics, availability, diversity, production and marketing conditions, guarantee of fair remuneration to producers, innovative nature, eco-friendliness, development of direct supply of agricultural products, vocational integration of disadvantaged groups, biodiversity, and animal welfare.

Delivery This includes the following aspects of the bid: delivery times, delivery conditions, customer service, technical support, supply reliability, interoperability, and operational characteristics.

Staff This category focuses on personnel-related aspects of the bid: organization, and professional qualifications and experience.

In addition to these categories, *ad hoc* criteria can be used, but they must be justified by the contract object or delivery conditions.

[9] http://www.marche-public.fr/Marches-publics/Definitions/Entrees/Criteres-choix-offres.htm

2 Presentation of the TED

In this section, we first describe the TED, which is our main data source (Section 2.1). We then turn to the data themselves and their structure (Section 2.2), before listing the issues that we detected (Section 2.3).

2.1 Tenders Electronic Daily

As mentioned before, the *Tenders Electronic Daily* (TED) is the online version of the supplement of the OJEU that is dedicated to the publication of the calls for tenders and award notices related to public procurement. Consequently, this site hosts documents related to all the public procurement contracts whose estimated cost is above the European Threshold (see Section 1). In addition, it can also host contracts below this threshold, but such publication is not compulsory.

2.1.1 Access and Content

There are two ways to access the content publicly hosted by the TED: by querying the database through an online API[10], or by downloading the data under the form of CSV files. Each such file covers a period of either one or ten years. Note that the API provides more information than the CSV files. For now, we use these files only though, as they seem to provide all the information we need in the context of DeCoMaP. These files are not directly stored on the TED, but rather on *data.europa.eu*[11], a website dedicated to hosting the EU open data. It offers two types of notices: contract notices and contract award notices[12].

A *contract notice* (CN) is a document that provides information about an upcoming contract, possibly divided in several lots. A *contract award notice* (CAN) provides information on the result of the selection process. This process can be split in several parts, each one constituting a *contract award* (CA). A contract award notice gathers information regarding the contract itself, but also the contract awards. It consequently contains information about the buyer (fields starting with `CAE`) and about the winner (fields starting with `WIN`). This is enough to connect a contract directly to a winner and a buyer, which is why we only focus on contract award notices for now. The TED offers CSV files listing all award notices starting from 2006. However, according to the documentation available on `data.europa.eu` [6], award notices published in the TED between 2006 and 2009 are both less complete and less reliable. This documentation also describes the content of the different fields in the database.

2.1.2 Versions

The format used by the OJEU to represent the contract notices and contract award notices changes through time to fit the evolution of laws and rules. In TED, notices are represented as XML files, whose structure is specified through an XML schema using the XSD dialect. This schema changed over time, in accordance with the modifications underwent by the notice format and structure.

The CSV files available on the `data.europa.eu` website were created in 2016, after the last major change in the notices format, which took place in 2014. Therefore, these data are represented using the most recent format, which is version 2. A specific field `XSD_VERSION` explicitly states the version number of the XML schema associated with each notice or contract notice, according to the CSV used.

The notices available in the TED rely on 5 distinct minor versions of the XML schema:

[10] https://ted.europa.eu/api/v2.0/notices/search
[11] https://data.europa.eu/
[12] https://data.europa.eu/data/datasets/ted-csv?locale=en

- Versions 2.0.5, 2.0.6 and 2.0.7: between 2006 and 2009;
- Version 2.0.8: since 2009;
- Version 2.0.9: since 2014.

Most of the notices in the TED use versions 2.0.8 or 2.0.9.

Version 2.0.8 This version is still used for some types of forms, especially those related to the defense and security sector. It is compliant with directive 2009/81/EC[13].

Version 2.0.9 This version is compliant with the directives 2014/23/EU[14], 2014/24/EU[15], and 2014/24/EU[16]. It essentially brings two main changes to the data structure.

First, it adds 14 new variables (the complete list can be found in the *Version* column of the tables provided in Appendix D). Among them, two are mandatory, and particularly important for us:

- `WIN_NATIONALID`: national identification number of the winner.
- `CRIT_PRICE_WEIGHT`: weight associated with the price criterion.

They are important because they allow us to build various forms of networks based on these tabular data. However, due to their late inclusion, they are not filled in notices relying on older versions of the XML schema. As we will see later, working with these notices requires some work to complement the missing information.

Second, some fields previously describing the whole contract have been moved lower, to the level of the single lot. These include the fields `ID_LOT`,`ADDITIONAL_CPV`, `B_VARIANTS`, `B_OPTIONS`, `B_EU_FUNDS`, `DURATION`, `CONTRACT_START`, `CONTRACT_COMPLETION`, which are fields describing the lots. This allows to provide different information for each lot. By comparison, before this change, all the lots had to share the same information.

2.2 Dataset Description

The TED gives access to the award notice of each EU public procurement contract above the European Threshold since 2006, which corresponds to $2{,}585{,}752$ award notices and $8{,}493{,}071$ lots. Data quality was improved in 2009 and the CPV typology was also revised in 2008, which is why we focus on the 2010–2020 period, as it allows us to deal with a stable set of fields.

The notices are published, on the one hand, by the (then) 28 EU member states, and on the other hand, by five affiliated countries willing to access the single market: Iceland, Liechtenstein, the Former Yugoslav Republic of Macedonia, Norway and Switzerland. For the period of interest, the TED contains $2{,}106{,}606$ award notices, corresponding to $7{,}169{,}070$ lots. The most represented countries are Poland and France, as shown in Figure 1. Romania is third, mainly due to an increasing number of published notices in recent years. Surprisingly, Germany and the UK come only fourth and fifth, respectively. This is probably due to different habits regarding the handling of public procurement, in particular proposing larger lots, and consequently, fewer of them. Table 34 (Appendix C) provides the exact numbers used to draw Figures 1 and 2. In the context of DeCoMaP though, we focus only on the French contracts, amounting to $410{,}283$ award notices (19.5%), and $1{,}380{,}965$ lots (19.2%).

The whole TED dataset takes the form of a single logical table. This table is broken down into several CSV files, each one representing a single year. In this table, each row represents a specific *lot*, which is described through 75 distinct fields. We distinguish four categories of fields: *Notice Metadata* (Section 2.2.1); *Agent Information* (Section 2.2.2); *Lot Information* (Section 2.2.3) and *Award Information* (Section 2.2.4). The interested reader will find the comprehensive list in Appendix D. In the rest of this section, we only focus on the

[13] https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1480931705809&uri=CELEX:32009L0081
[14] https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1480931610496&uri=CELEX:32014L0023
[15] https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1480931610496&uri=CELEX:32014L0024
[16] https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1480931610496&uri=CELEX:32014L0025

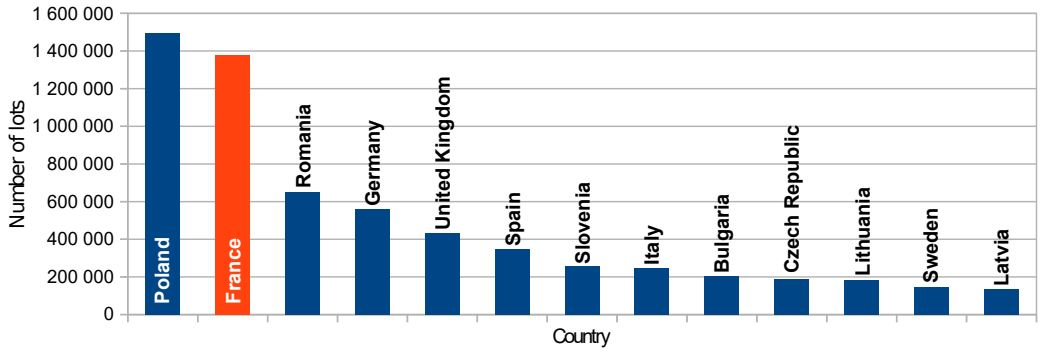


Figure 1. Number of lots published on the TED between 2010 and 2020, for countries with more than 100,000 lots.

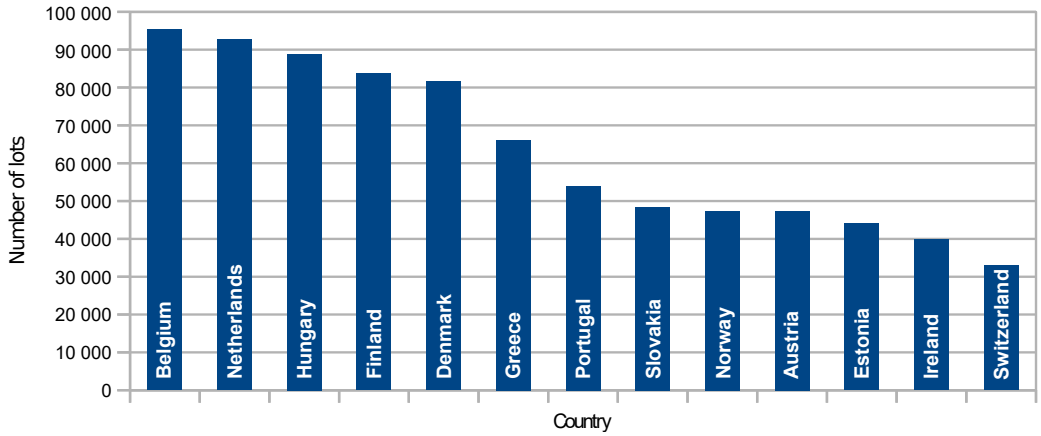


Figure 2. Number of lots published on the TED between 2010 and 2020, for countries with fewer than 100,000 lots.

fields which are the most relevant to our work, i.e. that can be used to build our network or help characterize fraud in public procurement.

2.2.1 Notice Metadata

This category gathers fields providing general information regarding the award notice. It includes:

- `ID_NOTICE_CAN`: unique identifier of the notice.
- `ID_LOT`: unique identifier of the concerned lot.
- `TED_NOTICE_URL`: URL of the notice on the TED website (page available only during 5 five years after publication).
- `YEAR`: year of publication of the call for tender notice.
- `CANCELLED`: whether the contract was canceled, and therefore not awarded.
- `CORRECTIONS`: number of corrections underwent by the contract after the publication of the call for tender.
- `INFO_ON_NON_AWARD`: if the contract was *not* awarded, indicates the reason why.

An open call for tender can be amended, and a contract can be canceled even after the end of the acceptance period: see Section 1.1 for more details.

2.2.2 Agent Information

The notion of *economic agent* refers to both the buyer and winner that enter a contract at the end of the awarding process. In the TED dataset, the buyer is called CAE, which stands

for *Contracting Authority or Entity*, and the candidate which is awarded the contract is called the *winner*.

The fields describing these agents in the dataset are the following:

- `NATIONALID`: unique identifier of the agent, specific to the concerned country.
- `NAME`: name of the agent.
- `ADDRESS`: postal address, composed of the street number, street type and street name.
- `POSTAL_CODE`: zipcode of the agent.
- `TOWN`: city of the agent.

These fields are similar for both buyers and winners, except that they are prefixed by `CAE_` for the former and `WIN_` for the latter. Public authorities have an additional field `CAE_TYPE` representing their type of public authority.

In the case of French contracts, the national identifier is the SIRET (cf. Section 2.3.7).

2.2.3 Lot Information

Fields from this category provide information regarding the lot sold through the considered contract. They include:

- `CPV`: main common procurement vocabulary code of the lot (cf. Section 1.1).
- `TYPE_OF_CONTRACT`: object of the lot, which can be *works*, *supplies* or *services*.
- `TOP_TYPE`: type of procedure used for the award.
- `CRIT_PRICE_WEIGHT`: importance weight given to the price criterion.
- `CRIT_CRITERIA`: list of criteria, except the price.
- `CRIT_WEIGHTS`: importance weights given to these criteria.
- `B_ON_BEHALF`: whether the contract involves several buyers buying together.
- `B_INVOLVES_JOINT_PROCUREMENT`: whether the contract is a joint procurement.
- `B_FRA_AGREEMENT`: whether the contract concerns a framework agreement.
- `B_GPA`: whether the contract is under the Government Procurement Agreement (GPA), which aims at regulating public procurement worldwide.
- `B_ACCELERATED`: whether the award procedure of the contract was accelerated.
- `OUT_OF_DIRECTIVES`: whether the award notice is published even without a contract notice.

As explained in Section 1.4, the award criteria are aspects of the supplier's bid that are considered by the CAE to select the winner. A contract lot may contain several award criteria, so each criterion has a weight to measure its relative importance. The TED assumes that price is always a criterion, and consequently has a dedicated field to represent its weight (`CRIT_PRICE_WEIGHT`). The other criteria are represented jointly, using two fields: one lists their names (`CRIT_CRITERIA`) and the other their weights (`CRIT_WEIGHTS`).

2.2.4 Award Information

Fields from this category provide information regarding the awarding process:

- `AWARD_VALUE_EURO_FIN_1`: value of the contract as eventually agreed by the buyer and winner.
- `NUMBER_OFFERS`: number of bids received by the buyer.
- `DT_AWARD`: date of the contract award.
- `B_CONTRACTOR_SME`: whether the contract was awarded to an SME (*Small and Medium-sized Enterprises*, i.e. fewer than 250 employees).
- `NUMBER_TENDERS_SME`: number of tenders received from SMEs.
- `B_SUBCONTRACTED`: whether the contract was subcontracted.

2.3 Detected Problems

The literature shows that, generally speaking, the TED data has a number of issues [1, 3, 4]. We performed a thorough analysis of the French data, which confirmed these and allowed identifying other problems. This section aims at summarizing them so that we can propose some appropriate solutions later on.

2.3.1 Missing Notices

As explained in Section 2.1, the TED contains both contract notices and contract award notices. Technically, the connection between them is done through the `FUTURE_CAN_ID` field contained in the CN, that is supposed to refer to the `ID_NOTICE_CAN` field of its matching CAN. However, we observe that this connection is broken in a number of cases.

On the one hand, the `FUTURE_CAN_ID` field is empty in certain CNs, or contains an identifier that does not correspond to any existing CAN. In this case, we do not have access to any information about the results of the awarding procedure, or the identity of the winner, if there is any. On the other hand, certain CANs do not appear in the `FUTURE_CAN_ID` of any CN. This prevents from accessing some information stored in the CN only, such as advertisement duration and contract duration. In both cases, depending on the specific procedure and situation, such orphan notices can be normal (i.e. the counterpart notice *should not* be in the database), or an issue (it *should* be present).

These problems were also discussed by Csáki *et al.* [3, 4] when assessing the whole TED dataset for the 2009–2015 period.

Orphan CNs For CNs, the main source of confusion are unsuccessful procedures, which the CAE can advertise on the TED, *if it wants to*. Before version 2.0.9 of the notice XML schema (i.e. before 2014, see Section 2.1.2), there was no proper way of doing so. As explained in a European Commission report [5], and as confirmed by our observations, in practice, data clerks either published a modification notice indicating a cancelation; or an award notice indicating *Infructueux* (French for unsuccessful), or some synonym expression, as the winner. Because advertising an unsuccessful procedure is not compulsory, many clerks also simply let their CN without a matching CAN. Version 2.0.9 introduced a new field `PROCUREMENT_UNSUCCESSFUL` in the CAN, allowing to properly publish this information. Some clerks took advantage of this addition, however many also just stuck to their previous practices. It is worth mentioning that if a CAE re-tender a previously withdrawn contract, it must indicate this in the new CN. However, we did not find any such situation in the dataset.

Type of notice	Present	Correct	Missing	Total
Contract Notices (CNs)	284,498	7,650	186,607	478,854
	59.41%	1.60%	38.97%	100.00%
Contract Award Notices (CANs)	286,244	47,760	76,279	410,283
	69.77%	11.64%	18.59%	100.00%

Table 1. Assessment of the connection between contract notices and contract award notices. The first row shows the numbers of present and missing CANs when considering the CNs, whereas the second row shows the numbers of present and missing CNs when considering the CANs.

The first row of Table 1 represents the different possible situations for CNs, as well as their frequency in the dataset. *Present* denotes cases of CNs that have a matching CAN that is effectively present in the database. These include unsuccessful procedures advertised explicitly through a CAN (therefore, after 2014). *Correct* means that either no matching CAN identifier is indicated or there is one but the CAN itself is missing, and there is a proper justification for this. This is the case when the acceptance period was not over at the time

the dataset was extracted, i.e. the call for tenders was still going on, so no decision could be made at this time (thus no matching CAN). This also includes unsuccessful procedures advertised through a modification notice. Finally, *Missing* denotes the rest of the notices, i.e. the cases where the CAN identifier or the CAN itself is missing, without any justification. This includes unsuccessful procedures that were not advertised at all, which is legally acceptable, but also possibly regular procedure without CAN, which is not. These represent no less than 38.97% of the CNs, but we have no way to distinguish them at this stage.

Orphan CANs Like for the CNs, a CAN missing its matching notice can correspond to a legally acceptable situation: certain procurement procedures do not require publishing a contract notice (AWP, NOC or NOP). The same is true for contracts and purchase orders issued within a framework agreement, and for the renewal of framework agreements themselves. Finally, it is also possible that the CN was published before the period covered by the dataset. But the rest of the time, a missing CN constitutes a data issue.

The second row of Table 1 shows how many CANs are concerned by the different possible situations in the database. *Present* denotes cases of CANs for which we could find a matching CN. *Correct* means that the dataset contains no matching notice, but that this can be explained, either by the selected procurement procedure or by an early publication date, as explained above. Finally, *Missing* denotes the rest of the cases, where the matching notice is missing when it should be present in the dataset. These represent approximately 19% of the CANs, which is much fewer than for CNs, but still very high. Like for the CNs, these number includes both errors and situations that are actually legal. Indeed, a framework agreement CAN without a matching CN may be either an initial contract, in which case the CN should be present, or a renewal, in which case this is not compulsory. But without the CN, it is not possible to distinguish these two situations, as this information (new agreement vs. renewal) is stored there.

2.3.2 Joint Agent Description

As explained in Section 1, a public procurement *lot* can be awarded by several CAEs, and to several winners. However, in the TED dataset, each row represents a single lot, and all CAEs and winners are indicated jointly in their respective fields. Here is an example of lot involving several CAEs:

```
ID_NOTICE_CAN   CAE_NAME
      2018338   Centre hospitalier d'Arras---Centre hospitalier du Ternois
```

And here is an example of lot involving several winners:

```
ID_NOTICE_CAN   WIN_NAME
   2015283576   Montaigne --- BRGC
```

Table 2 shows the distributions of the lots according to the number of CAEs and winners, for the whole European Union and for France in particular. It appears that even if there is only a single CAE and a single winner in the overwhelming majority of lots, the number of lots with several agents is still significant, and should be handled properly.

In order to take advantage of these data, we need a separate representation of these agents, since we want to connect them afterwards. Therefore, we must split these values. As illustrated in the previous examples, the standard TED separator between two different agents is a triple hyphen `---`. Ideally, we should always get a string of the form `Agent A --- Agent B`.

However, this is not always the case, and some data entry clerks (or *clerks* for short in the rest of the document) adopt other ways to indicate a multiplicity of entities, using for instance a slash (`/`):

Number of agents per lot	Number of lots European Union Buyers	European Union Winners	France Only Buyers	France Only Winners
1	7,041,375	6,937,625	1,346,708	1,343,522
2	66,822	136,125	21,474	22,573
3	14,224	52,240	3,948	7,864
≥ 4	46,649	43,080	8,835	7,006

Table 2. Number of agents per lot for European and French contracts, based on the official triple hyphen (---) separator string.

```
ID_NOTICE_CAN   WIN_NAME
      2010358   Scape architecture / Treuttel Garcias / Lan architecture
```

2.3.3 Name Inconsistency

As explained in Section 2.2.2, for each lot, TED provides the names of involved agents. However, this field is not normalized, in the sense that one agent can be named using different strings. This is an issue, because this makes agent identification more difficult. We can separate this problem into three sub-problems: inconsistencies in the use of typography, occurrences of different proper nouns to refer to the same agent, and inclusion of irrelevant information in the name field.

Typographic Inconsistency Names, like other string fields in TED, are not normalized: diacritics, and punctuation signs are not used consistently. This makes it impossible to directly perform exact matching between these strings.

```
ID_NOTICE_CAN   CAE_NAME
      2010334   Commune du Grau du Roi

ID_NOTICE_CAN   CAE_NAME
      2010334   Commune du Grau-du-Roi
```

Multiple Proper Nouns Sometimes, an agent can be named using different strings in the TED. A common case is the non-systematic use of acronyms, for instance:

```
ID_NOTICE_CAN   CAE_NAME
      2010334   CEA

ID_NOTICE_CAN   CAE_NAME
      2010334   Commissariat à l'énergie atomique et aux énergies alternatives
```

The French Atomic Energy Commission, mentioned above, appears 1,264 times under the acronym form of its name (`CEA`), and 1,021 under its full name.

Name Pollution Sometimes, the name field also contains additional information related to the physical location of the agent (ex. building number), or its role in a larger structure (ex. internal department). Here is an example of the latter type:

```
ID_NOTICE_CAN   CAE_NAME
   2013265707   Réseau ferré de France - direction régionale Centre-Limousin
```

This information is not related to the agent's name itself, and makes it harder to perform a proper comparison. As an example, the words `direction` and `service`, which often refer to some internal departments, appear in 95,185 and 60,854 of addresses, respectively.

2.3.4 Address Inconsistency

We detected four types of problems with the addresses. First, there is a normalization problem as for the agent's names (cf. Section 3.4.2): typography is not used consistently. Second, the TED confuses several types of addresses (postal, geographic, and geopostal). Third, the fields used to store addresses mix various aspects in an inconsistent way. Fourth, certain address fields sometimes contain irrelevant information.

Typographic Inconsistency The address and town fields use hyphens, diacritics and abbreviations in an inconsistent way. Here are two towns differing only by the use of an abbreviation:

```
ID_NOTICE_CAN   CAE_TOWN
      2010142   Saint-Julien-en-Genevois

ID_NOTICE_CAN   CAE_TOWN
      2010142   St-Julien-en-Genevois
```

And here is another town written with and without hyphens:

```
ID_NOTICE_CAN   CAE_TOWN
   2013265407   Valence-d'Agen

ID_NOTICE_CAN   CAE_TOWN
   2013265407   Valence d'Agen
```

We detect $602{,}470$ agent occurrences whose names contain hyphens.

Type Confusion A database can contain three possible types of addresses. A *geographic* address indicates the physical location using information such as building number, street number, street type, city, and country. A *postal* address is designed for the purpose of mail delivery: it contains only the information used by the postal service for delivery purposes, e.g. zipcode, post office box number. Finally, a *geopostal* address contains both types of information.

In the TED, all three types of addresses appear. Here is an example of geographical address:

```
ID_NOTICE_CAN   CAE_ADDRESS
   2017373033   13 place Vendôme
```

Here is an example of a postal address:

```
ID_NOTICE_CAN   CAE_ADDRESS
   2017372306   CS 20100
```

Here is an example of a geopostal address:

```
ID_NOTICE_CAN   CAE_ADDRESS
   2017373096   place Maurice Mollard; BP 348
```

For example, `CS`, `CEDEX` and `BP`, which are three indicators related to postal and geopostal addresses, appear in $217{,}078$, $26{,}135$ and $469{,}008$ addresses, respectively.

The mixed occurrence of all three address types (geographical, postal, and geopostal) makes it difficult to compare addresses within the TED, or even to external sources, because of this lack of consistency.

Monolithic Address The `xxx_ADDRESS` field combines several parts of a geographic address, which are usually considered separately in standard databases: street number, street type and street name. For instance, in the previous example, the field value is `13 place Vendôme`, which combines all three address parts.

The fact that these parts are combined in the TED makes it difficult to compare its addresses to those coming from other sources, as we will see later.

Address Pollution We call address pollution the presence of irrelevant information in certain address fields, in particular `xxx_TOWN`. For example, for certain agents, the CEDEX code (*Courrier d'Entreprise à Distribution EXceptionnelle* – an accelerated postal service for companies) is specified after the city name:

```
ID_NOTICE_CAN   CAE_TOWN
      2010195   Grenoble Cedex 9
```

This word `CEDEX` appears in the town field of $813{,}421$ agent occurrences.

Sometimes, the district or locality is indicated in the same field, for instance:

```
ID_NOTICE_CAN   CAE_TOWN
   2017373089   Paris La Défense
```

As for the previous error types, these mistakes make it difficult to compare addresses, both within the TED or from external sources.

2.3.5 Unconstrained Criterion Description

As explained in Section 2.2.3, the TED lists the award criteria and their weights. This concerns $1{,}056{,}100$ lots, i.e. 76% of the dataset. However, the way this information is modeled makes it difficult to use, as clerks do not always adopt the same convention to fill the fields. This is an issue for us, because award criteria are likely to constitute a discriminant information in the context of corruption or fraud prediction [9].

First, as for the *Joint Agent Description* issue, multiple criteria and weights are sometimes shoved into the same field. The string that appears to be the standard TED separator, as before, seems to be the triple hyphen `---`: it appears in the criterion name or weight fields in $38{,}107$ lots (3%). However, it is not used systematically: data entry clerks alternatively put a slash `/`, a semicolon `;`, or other characters. Here is an example of inappropriate separator for three criteria (technical value, delivery time, and price):

```
 ID_LOT   CRIT_CRITERIA
2013466   VALEUR TECHNIQUE/DELAI DE LIVRAISON/PRIX
```

Second, the fields containing criterion names allow the data clerk to type free text, by opposition to the selection of certain criteria in a closed, predefined list. Consequently, the names present in the TED widely differ, making it difficult to infer that two strings actually represent the same criterion. Here is an example of two strings with representing the same criterion in two very different ways:

```
 ID_LOT   CRIT_CRITERIA
1096365   ENGAGEMENT DU PRESTATAIRE EN TERME DE DEVELOPPEMENT DURABLE
1097818   PROTECTION DE L'ENVIRONNEMENT
```

Both can be categorized as environmental criteria, but the first one translates to *Supplier hired depending on sustainable development* (with a typo, on top of that) whereas the second is *Protection of the environment*.

Third, the price weight is sometimes mixed with the weight of the other criteria, like for instance:

```
 ID_LOT   CRIT_CRITERIA             CRIT_WEIGHTS
2010169   PRIX---VALEUR TECHNIQUE   40---60
```

Here, the price criterion is listed together with the technical value, and so are its weight. This type of issue affects $798{,}794$ lots (58%).

Fourth, the weights associated to the criteria are not normalized, i.e. the bounds are not fixed, and they can sum to any value:

```
ID_LOT   CRIT_CRITERIA                                CRIT_WEIGHTS
672086   CONFORMITE AU CAHIER DES CHARGES---          4---3---3
         VALEUR TECHNIQUE DE L'OFFRE---PRIX
```

In the above example, the weights of the vocational integration and technical value criteria sum to 10.

To the best of our knowledge, there is no official recommendation regarding the bounds of these weights. However, in practice, we observe that most weights are expressed as percents. We thus normalize the criterion weights associated to a lot in order to make them sum to 100.

There are $176{,}504$ lots (13%) that do not implement this rule. In order to make these values comparable from one lot to the other, we need to normalize them over the whole dataset.

Fifth, sometimes the criterion names and weights are put together in the same field, for instance:

```
ID_LOT   CRIT_CRITERIA
672086   QUALITE 50 %---PRIX 30 %---PRESTATION 20 %
```

Here, we have three criteria: quality, price, and delivery, whose respective weights are 50, 30 and 20. This issue affects $23{,}918$ lots (2%).

Sixth, each TED row should only contain the information related to the considered lot. However, it happens that the criteria of all the lots constituting a given contract are described in the same row, for instance:

```
ID_NOTICE_CAN   CRIT_CRITERIA
      2010227   Evaluation financière (lots 1 - 2 - 8- 9 et 10)
                ---Valeur technique (lot 1- 2 - 8- 9 et 10)
                ---Prestations de service (lot 1- 2 - 8- 9 et 10)
                ---Evaluation financière (lots 3 - 4)
                ---Valeur technique (lots 3 - 4)
                ---Prestations Evaluation financière (lots 1 - 2 - 8- 9 et 10)
                ---Valeur technique (lot 1- 2 - 8- 9 et 10)
                ---Prestations de service (lot 1- 2 - 8- 9 et 10)
                ---Evaluation financière (lots 3 - 4)
                ---Valeur technique (lots 3 - 4)
                ---Prestations de service (lots 3 - 4)
                ---Evaluation financière (lots 5 - 6)
                ---Valeur technique (lots 5 - 6)
                ---Prestations de service (lots 5 - 6)
                ---Evaluation financière (lot 7)
                ---Prestations de service (lot 7)de service (lots 3 - 4)
                ---Evaluation financière (lots 5 - 6)
                ---Valeur technique (lots 5 - 6)
                ---Prestations de service (lots 5 - 6)
                ---Evaluation financière (lot 7)
                ---Prestations de service (lot 7)
```

Here, the contract involves two lots, with different criteria. All of them are listed in both rows describing these two lots, instead of only listing the criteria of each concerned lot at once.

The first lot uses price as its sole criterion, whereas for the second lot there are two criteria: price and candidate involvement. This issue affects $39{,}234$ lots (3%).

In order to use this data, we need to identify which criterion applies to which lot, and this requires solving all these issues.

2.3.6 Incorrect Contract Prices

Another important issue concerns the price associated to each contract in the TED. This information is likely to be stored in no fewer than 9 distinct fields [6]. The first three are located in the contract notices. Three fields of identical names appear in the award notices, but they have a slightly different meaning. The last three fields are only present in the award notices. Here is the complete list:

- Contract notices (CN):
  - `VALUE_EURO`: contract value as originally estimated by the buyer.
  - `VALUE_EURO_FIN_1`: same value as in `VALUE_EURO`, or framework value if this field is empty.
  - `VALUE_EURO_FIN_2`: generally the same value as `VALUE_EURO_FIN_1`, but possibly manually corrected by EU services.
- Contract award notices (CAN):
  - `VALUE_EURO`: for a contract, total value of the winning bid(s), or of the lower bid(s).
  - `VALUE_EURO_FIN_1`: same as in `VALUE_EURO`, or an automatic estimation if this field is empty.
  - `VALUE_EURO_FIN_2`: generally the same value as `VALUE_EURO_FIN_1`, but possibly manually corrected by EU services.
  - `AWARD_EST_VALUE_EURO`: for a contract award, estimated value of the winning bid.
  - `AWARD_VALUE_EURO`: for a contract award, effective value of the winning bid, or lowest bid if this value is missing.
  - `AWARD_VALUE_EURO_FIN_1`: value provided if field `AWARD_VALUE_EURO` is empty. It is estimated based on other fields.

All these are pre-tax values, expressed in Euros.

All the fields starting with `VALUE_` describe the whole contract, whereas those starting with `AWARD_` concern a single contract award. Consequently, for a given contract, the sum of all `AWARD_EST_VALUE_EURO` should equal field `VALUE_EURO` from the CN; and the sum of all `AWARD_VALUE_EURO` should equal field `VALUE_EURO` in the CAN.

...

We detect mainly two problems with these price fields: the value can just be plainly missing, or it can be present but incorrect.

Missing values In a number of cases, the price is simply missing, for instance:

```
ID_NOTICE_CAN   AWARD_VALUE_EURO
      2010169   NULL
```

Table 3 indicates the missing rate for all the price-related fields in the award notices. For each such field, it shows the number of lots without any information (column *Lots*), and the corresponding proportion (column *%*). Note that the completion rates of all fields are available in Appendix C.1.

It appears that the price information is missing in most of the lots, which makes it completely impossible to leverage this information. The same remark was made in a communication of the EU [8] for the whole TED data over period 2014–2016: "Although the TED data is of reasonably good quality regarding the number of contracts recorded, some challenges remain regarding the reporting of contract values. These lead to either a lack of an actual contract value in the database or the use of an arbitrary low value."

Field	Lots	%
VALUE_EURO	654,493	47.4
VALUE_EURO_FIN_1	948,767	68.8
VALUE_EURO_FIN_2	948,767	68.8

Field	Lots	%
AWARD_EST_VALUE_EURO	189,613	13.8
AWARD_VALUE_EURO	815,928	59.1
AWARD_VALUE_EURO_FIN_1	954,642	69.2

Table 3. Missing information in the price-related fields.

Aberrant Values Moreover, even when a price field is filled, the provided information is not always reliable. In certain cases, the value is simply aberrant, for instance there are 16,662 lots (1.2%) whose AWARD_VALUE_EURO field indicates only €1. This is typically the case when a contract involves many identical items, and we assume that the data clerk incorrectly indicates the unitary cost of one item instead of the total price of the lot. The large number of contract with such an abnormally low price was also mentioned in the EU communication quoted above [8].

On the other end of the spectrum, a report from the European Commission [1, p.5, 9 & 26] notices "impossibly high figures for the value of contract awards". This is a general observation made for the whole TED data, but it holds for French notices.

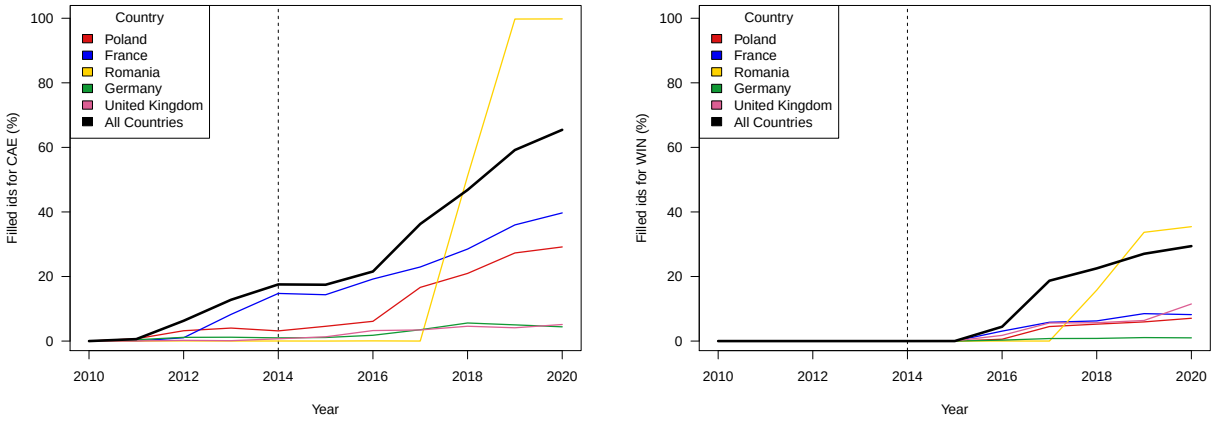
This might be related to another type of dubious prices, constituted of the same digit repeated many times, as in the following example:

```
ID_NOTICE_CAN    AWARD_VALUE_EURO
   2016176878    9999999999
```

...

2.3.7 Missing Agent Identifiers

The TED contains a unique identifier to identify each economic agent. This number is national, and can be different for each country. In the case of France, it is the SIRET (*Système Informatique pour le Répertoire des Entreprises sur le Territoire* – Computer system for the national register of companies), which is a 14-digit number representing a specific facility in France (see Section 4.1 for more details about this).



(a) CAE (buyer) filled identifiers.

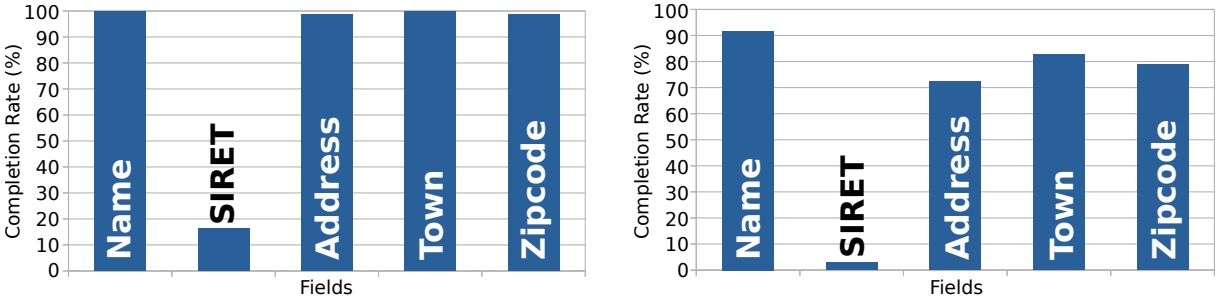
(b) Winner (supplier) filled identifiers.

Figure 3. Evolution of the completion rate for the identifier field over the considered period (2010–2020) for the 5 countries publishing the most notices, and for the whole dataset (All). The dotted line materializes the change in notice version (2014).

In theory, the buyer's national identifier was already required in version 2.0.8 of the notices, i.e. before 2014, whereas the winner's national identifier is required since version 2.0.9, i.e. after 2014 [5]. However, this is not true in practice, as shown by Figure 3, which exhibits

the evolution of the completion rate regarding the identifier field, for the main providers of notices in the TED, as well as the whole dataset. The completion rate is far from perfect, especially for France, which is of particular interest to us. The trend shows an increase in the number of filled identifiers though, starting before 2014 for buyers and after 2014 for winners. This difference is likely due to the change in regulation mentioned before. The detail of the completion rate for the identifier field, regarding all countries described in the TED dataset, is available in Appendix C.3.

We now focus on France, and it appears that data entry clerks rarely fill the SIRET, even in recently published notices. Figure 4 shows the completion rate for a selection of TED fields related to economic agents. The y axis shows the proportion of lots for which the field is filled. The figure represents separately the CAEs and the winners. The detail of the completion rates for the other fields is available in Table 33 (Appendix C.2). It appears that the SIRET is filled in only 16.4% of the lots for the buyers, and 2.9% for the winners. This crucial information is therefore extremely scarce in the TED data.



(a) CAE (buyer) fields.

(b) Winner (supplier) fields.

Figure 4. Completion rate for fields related to the identification of economic agents.

This means that we are not sure of the exact identity of the buyer or winner most of the time. Eventually, our objective is to extract various types of graphs from the TED dataset, considering CAEs and winners as nodes. Therefore, it is important to correctly identify all instances of the same economic agent. Otherwise, it is likely that the same agent will be represented as several distinct nodes, which would affect the graph structure.

To solve this problem, we need to leverage the different aspects of the information we have about the agents: name, location, and activity domain. Their location is described in fields address, city and zipcode. However, clerks do not always fill all these fields in public procurement notices, as shown in Figure 4. They are filled most of the time for CAEs, but only approximately 75% of the time for winners. The general better level of completion for CAEs compared to winners could be due to the fact that CAEs complement the contract award notice, and therefore better fill their parts. More generally, regarding the whole TED dataset, a report of the European Commission [1] stresses that "During the period 2009-2015, 15 % of the TED mandatory fields were empty".

The activity domain is more difficult to handle, because in the TED it is not described at the level of the agent, but rather at the level of the lot. The `CPV` field (Common Procurement Vocabulary) contains the main CPV code associated with a lot. It gives one of the main characteristic of the contract, and is *always* filled. As mentioned in Section 1.1, each of these codes is defined as a part of a larger typology describing all subjects handled in public procurement. Although this is a lot field, we can still use it to obtain additional information on the winner, assuming that the winner's activity domain is related to this CPV code. However, in France, the activity domain of a company is represented by the `APE` code (*Activité Principale Exercée* – Main Pursued Activity). We did not find any correspondence between CPV and APE, so we had to create our own mapping.

2.4 Overview of the Proposed Method

Our approach to solve the problems identified in Section 2.3 is described in Figure 5. It contains 4 steps that we summarize here, and describe in detail in the rest of this document.

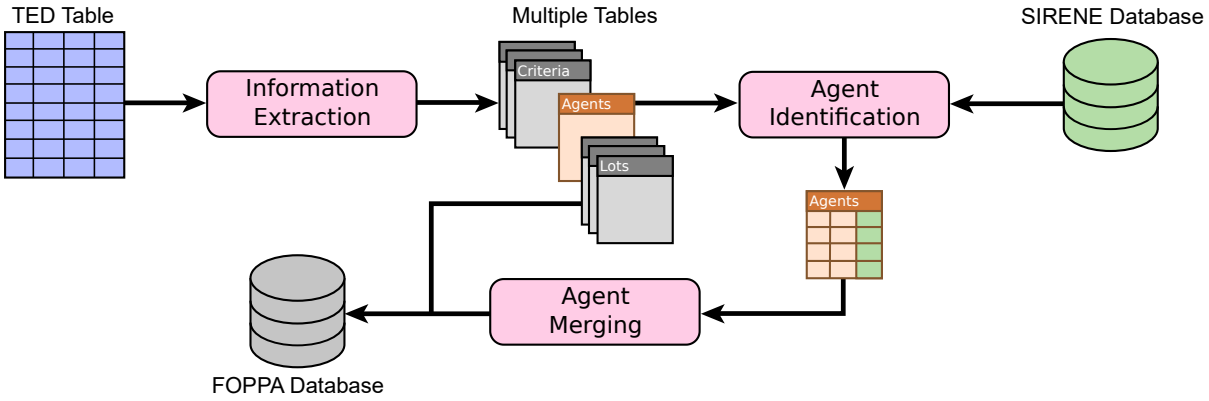


Figure 5. Overview of our proposed method to clean the TED data before graph extraction.

Information Extraction As mentioned in Section 2.1, the TED dataset is constituted of a single table broken down into several yearly CSV files, where each row represents an individual lot, with several potential CAEs and winners involved. To ease the data verification and future use, our first step consists in splitting this table into three separate new tables representing the lots, the award criteria appearing in their description, and the concerned agents.

During this step, we solve the *Joint Agent Description* (Section 2.3.2) and *Unconstrained Criterion Description* (Section 2.3.5) problems. We also tackle certain aspects of the *Address Inconsistency* (Section 2.3.4) and *Name Inconsistency* (Section 2.3.3) problems. We describe this step in Section 3.

Agent Identification In this step, we tackle the remaining aspects of the *Address Inconsistency* (Section 2.3.4) and *Name Inconsistency* (Section 2.3.3) problems. We also start dealing with the *Missing Agent Identifiers* problem (Section 2.3.7).

The task that we call *siretization* consists in retrieving the missing SIRETs. For this purpose, in this identification step, we take advantage of SIRENE, an external database maintained by the French state, and listing all existing SIRETs ever. We describe this step in Section 4.

Cluster-Based Merging Our identification process is not able to find a reliable SIRET for all agents, because of missing or inaccurate data. The goal of this step is to deal with the remaining cases, therefore finishing solving the *Missing Agent Identifiers* problem (Section 2.3.7). We call an agent *identified* if it has a SIRET, and *siretized* if, more specifically, this SIRET was originally unknown in the TED data, and later retrieved through our process. An agent is considered *unidentified* when it has no SIRET.

We use a fuzzy matching library called `Dedupe`, in order to group similar agent instances thanks to their address and name. Based on this process, we can get two types of clusters. If a cluster contains only unidentified agent instances, then we can assume these are different forms of the same entity and merge them. If a cluster contains both identified and unidentified agents, then we can assume that the latter instances are instances of the former agents. We describe this step in Section 5.

3 Step 1: Database Initialization

The goal of this first step is to split the single original TED table into several separate tables, in order to ease both the cleaning and usage of the data. In Section 3.1, we describe the structure of our database. Then, we explain how we process the original TED data in order to split them and fill our database. In Section 3.2 we separate multiple criteria and their respective weights. In Section 3.4, we focus on addresses and agent names.

3.1 Database Structure

Our FOPPA database contains six tables, as described in Figure 6, and detailed in the rest of this section.

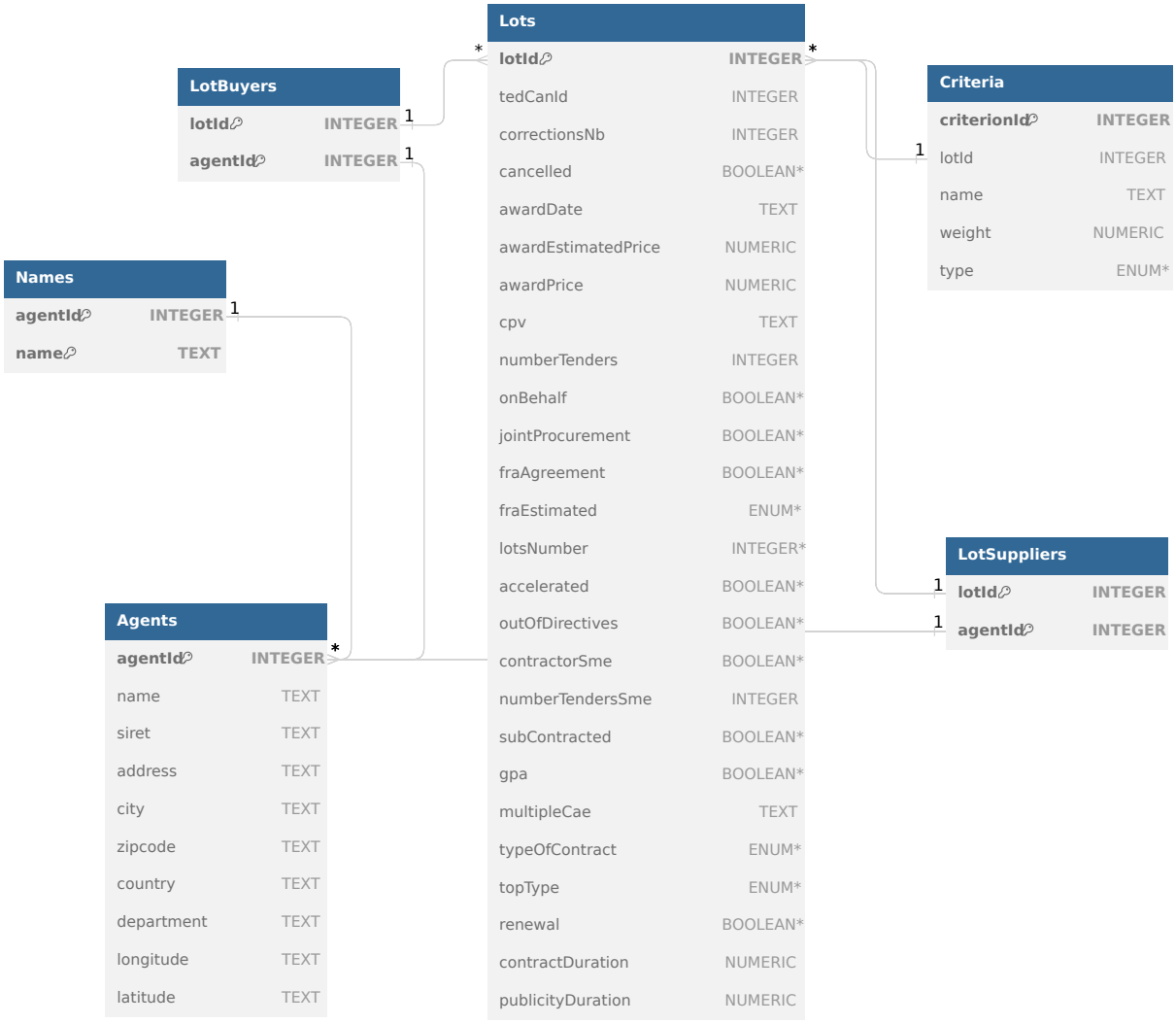


Figure 6. Structure of our database, shown as an Entity-Relation diagram. The types are theoretical, as in practice, those with stars are different in the SQL implementation.

We must point out that the data types indicated here are only *theoretical*, in the sense that they correspond to what one would expect in an ideal situation. But in practice, the `lotsNumbers` and `contractorSME` fields are not perfectly clean and contain values that do not fit the stricter constraints of these theoretical types, so we must use text, a more relaxed type, in the actual database.

Another limitation concerns enumerated and Boolean fields, as these types are not allowed in certain SQL variants. To make our database compatible with such systems, we

use text to represent the former, and integers (0 vs. 1) for the latter.

Table `Lots` The lots are the central information in our dataset. We represent them in a dedicated table, which contains the following fields:

- `lotId`: unique identifier of the lot.
- `tedCanId`: TED identifier of the contract award notice.
- `correctionsNb`: number of correction notices published for the lot.
- `cancelled`: Boolean value indicating whether the lot was cancelled.
- `awardDate`: date the lot was awarded.
- `awardEstimatedPrice`: estimation of the value of the lot, according to the *contract* notice.
- `awardPrice`: value of the lot in the *award* notice.
- `cpv`: main *Common Procurement Vocabulary* (CPV) code of the lot (cf. Section 1.1).
- `tenderNumber`: number of supplier offers for the lot.
- `onBehalf`: Boolean value indicating whether the lot involves several buyers buying together.
- `jointProcurement`: Boolean value indicating whether the lot involves a joint procurement.
- `fraAgreement`: Boolean value indicating whether the lot involves a framework agreement.
- `fraEstimated`: if appropriate, nature of the information suggesting that the lot involves a framework agreement.
- `lotsNumber`: number of lots in the contract award notice.
- `accelerated`: Boolean value indicating whether the procedure was accelerated.
- `outOfDirectives`: Boolean value indicating whether a CAN was published without CN.
- `contractorSme`: Boolean value indicating whether the buyer is an SME.
- `numberTendersSme`: number of SME offers for the lot.
- `subContracted`: Boolean value indicating whether the lot was subcontracted.
- `gpa`: Boolean value indicating whether the lot was associated to the Government Procurement Agreement (GPA).
- `multipleCae`: Boolean value indicating whether the CAN lists multiple contracting authorities.
- `typeOfContract`: type of the contract, one among:
  - `S`: Supplies;
  - `W`: Works;
  - `U`: Utilities.
- `topType`: Type of awarding procedure, one among (cf. Section 1):
  - `AWP`: award without prior publication of a contract notice;
  - `COD`: competitive dialogue;
  - `NOC/NOP`: negotiated without a prior call for competition;
  - `NIC/NIP`: negotiated with a call for competition;
  - `OPE`: open procedure;
  - `RES`: restricted procedure;
  - `INP`: innovative partnership.

These fields directly come from the TED fields described in Section 2.2.3. Due to the central position of the concept of lot in our dataset, this is by far the largest table of our database.

Table `Criteria` As explained in Section 2, the number of criteria used to award a lot is not predefined, and can range from one to any number. Therefore, there is a many-to-many relationship between lots and criteria. In our base, the concept of criterion is just an enumerated value, though, so there is no need for a specific table to represent the criteria themselves. Instead, we need an association table modeling the association between a lot and its criteria.

In table `Criteria`, each row associates a specific criterion to a specific lot. In addition to the `lotId`, which acts as a foreign key, the table contains 3 fields describing the criterion itself:

- `name`: name of the criterion.
- `weight`: normalized weight of the criterion, relative to all other criteria selected for the concerned lot. It is expressed as a percentage.
- `type`: type of the criterion, which can take 6 possible values:
  - `PRICE`: price;
  - `DELAY`: deadline;
  - `TECHNICAL`: technical terms;
  - `ENVIRONMENTAL`: environmental terms;
  - `SOCIAL`: social terms;
  - `OTHER`: other types of terms.

Criterion names are originally expressed as free text in the TED. In order to ease criterion-ealted their analysis, we add a new field `type`, that corresponds to the broad category of criteria.

Table `Agents` In order to avoid duplicating the information related to economic agents as in the CSV version of the TED dataset, and in order to ensure data consistency, we store agent-related information in a dedicated table. In this table `Agents`, each row represents a single and unique economic agent, which is described using the following fields:

- `agentId`: unique identifier of the agent, *in our database*.
- `name`: principal name of the agent.
- `siret`: SIRET of the agent, i.e. unique identifier of the agent in the TED database (supposedly).
- `address`: full address of the agent.
- `city`: city of the agent.
- `zipcode`: zipcode of the agent.
- `country`: country of the agent.
- `department`: French department of the agent, a code containing 2 or 3 characters.

These fields directly come from the TED fields described in Section 2.2.2. In addition, we insert some extra information found in the SIRENE database (see Section 4.1), and taking the form of the following additional fields:

- `longitude`: the longitude of the agent.
- `latitude`: the latitude of the agent.
- `legalcat`: the legal category of the agent.
- `activityDomain`: the main domain of activity of the agent, according to SIRENE.

The legal category is represented according to the official French typology[17], which includes 306 classes distributed over 3 hierarchical levels.

We retrieve the longitude and latitude coordinates from a geolocated version of SIRENE[18].

Table `Names` An agent can be associated with several names. We create a table in order to keep every name. This table `Names` contains two fields constituting a multiple key:

- `agentId`: identifier of the agent.
- `name`: one of the names of the agent.

Tables `LotBuyers` and `LotSuppliers` As explained in Section 2.2, there can be several economic agents acting as buyers and/or as winners for a single lot. Therefore, we have a many-to-many relationship here. But unlike with the criteria, this time both agents and lots require dedicated tables to store all their related information. We consequently need two specific association tables to connect each lot to the relevant buyers and winners.

[17] https://www.insee.fr/fr/information/2028129

[18] https://public.opendatasoft.com/explore/dataset/sirene_v3/

Each row in table `LotBuyers` models the involvement of a specific economic agent as buyer for a specific lot. Table `LotSuppliers` has the same role for winners. Both tables contain the same fields:

- `lotId`: identifier of the lot.
- `agentId`: identifier of the buyer or winner.

3.2 Criterion Processing

The goal of this processing is to fix the *Unconstrained Criterion Description* problem (Section 2.3.5). For this purpose, we perform the following operations.

No Criteria Specified Some lots have no associated criterion at all in the TED, which is not licit (cf. Section 1.4). We could complete with a price criterion with a 100% weight, as it is a compulsory criterion. However, that would result in representing identically a lack of information and a lot originally associated with a single price criterion.

We prefer to explicitly show the absence of any criterion in the FOPPA, by not associating the lot to any entry in our `Criteria` table. This way, the end users can decide by themselves how to manage the lots with no criteria, and possibly assign a default price criterion if they see fit.

Weight Cleaning Using a regex, we parse the text strings present in both TED weight fields (`CRIT_PRICE_WEIGHT` and `CRIT_WEIGHTS`), and remove everything that is not a number or the standard delimiter (`---`). The objective of this process is to remove all superfluous words. Here is an example showing the field before and after this process:

```
ID_NOTICE_CAN   CRIT_WEIGHTS
     20102608   0;45---0;35---0;2
     20102608   045---035---020
```

Criteria Splitting Once the weights are clean, we proceed with the separation of multiple criteria and the computation and/or normalization of their respective weights. There are several possible cases to handle, which we list here from the simplest to the most complicated, in terms of data processing.

If both the criteria name and price weight fields (`CRIT_CRITERIA` and `CRIT_PRICE_WEIGHT`) are empty (which should not happen, legally speaking), or if only the price weight field (`CRIT_PRICE_WEIGHT`) is filled, then there is nothing to do at all.

If both the criteria name and weight fields (`CRIT_CRITERIA` and `CRIT_WEIGHTS`) are filled, we assume that all weights are located in the weight column. We then look for separation patterns, such as the usual triple hyphens (`---`), but also the slash (`/`). Moreover, we check that the number of separators is the same in both fields.

If only the criteria name field (`CRIT_CRITERIA`) is filled, we look for the same separation patterns as before. In this case, we have identified other formattings used when filling this field, but these are too heterogeneous and each one appears very rarely. We consider these cases are not worth the effort.

Here is an example of such a separation, operated on both `CRIT_CRITERIA` and `CRIT_WEIGHTS` fields:

```
ID_NOTICE_CAN   CRIT_CRITERIA                          CRIT_WEIGHTS
      2010142   Prix---Valeur Technique---Delai       40---40---20
>>
ID_NOTICE_CAN   critName                               critWeight
      2010142   Prix                                  40
      2010142   Valeur Technique                      40
      2010142   Délai                                 20
```

Here is another example, this time when only `CRIT_CRITERIA` is filled:

```
ID_NOTICE_CAN   CRIT_CRITERIA                              CRIT_WEIGHTS
      2010220   Critères Technique (note sur 10) ---       -
                Critères Economiques (note sur 10)
>>
ID_NOTICE_CAN   critName                                   critWeight
      2010220   Critères Technique (note sur 10)          -
      2010220   Critères Economiques (note sur 10)        -
```

Weight Extraction When the `CRIT_CRITERIA` field is the only one filled, we cannot get the weights directly, so we extract them using a regex. We then check that the number of weights found is equal to the number of criteria. If it is not the case, we remove the inconsistent weights, for example zeros. Here is an example of weight extraction:

```
ID_NOTICE_CAN   critName                                   critWeight
      2010220   Critères Technique (note sur 10)          -
      2010220   Critères Economiques (note sur 10)        -
>>
ID_NOTICE_CAN   critName                                   critWeight
      2010220   Critères Technique (note sur 10)          10
      2010220   Critères Economiques (note sur 10)        10
```

Weight Normalization We then normalize the weights in order to get relative values for each criterion. We apply the following formula to the old weights w_i in order to find the new weights w_i':

$$w_i' = \frac{w_i \times 100}{\sum_i w_i}. \tag{1}$$

Thanks to this, the sum of the weights of the criteria for each lot is 100, which makes it possible to make comparisons.

Here is an example of such normalization:

```
ID_NOTICE_CAN   critName                                   critWeight
      2010220   Critères Technique (note sur 10)          10
      2010220   Critères Economiques (note sur 10)        10
>>
ID_NOTICE_CAN   critName                                   critWeight
      2010220   Critères Technique (note sur 10)          50
      2010220   Critères Economiques (note sur 10)        50
```

If only the price weight field (`CRIT_PRICE_WEIGHT`) is filled, then we insert `Price` as a criterion in our own table, with a weight of 100%.

Criteria Classification The criteria names that appears in the TED are not normalized, which means that they are very heterogeneous. This makes it very difficult to compare contracts. To solve this issue, we define coarser categories of criteria which we store in our database, in a specific field `critType`, in addition to the original (free text) criterion names.

These classes are:

- `PRIX` (price);
- `DELAI` (deadline);
- `TECHNIQUE` (technical terms);
- `ENVIRONNEMENT` (environmental terms);
- `SOCIAL` (social terms);
- `AUTRES` (others).

We use regex to find keywords, for example `TECHNIQUE` (i.e. *technical*) or `DELAI` (i.e. *delay*), and assign the corresponding class to the cluster.

Here is an example of this process:

```
ID_NOTICE_CAN   critName                              critWeight
      2010220   Critères Technique (note sur 10)     10
      2010220   Critères Economiques (note sur 10)   10
>>
ID_NOTICE_CAN   critName                              critWeight   critType
      2010220   Critères Technique (note sur 10)     10           TECHNICAL
      2010220   Critères Economiques (note sur 10)   10           PRICE
```

The first notice is categorized as `TECHNICAL` due to the occurrence of keyword `Technique`, whereas the second is categorized as `PRICE` due to `Economiques`.

3.3 Lot Processing

As explained in Section 2.3.1, a CN is generally connected to a CAN through its `FUTURE_CAN_ID` field. For each CAN, this allows us to retrieve its matching CN and extract some information which is absent from award notices but present in contract notices only. We add some of this extra information to our database, under the form of the following fields. Of course, when there is no matching CN, these remain empty.

Advertising Period Contract notices contain the `DT_DISPATCH` field, which represents the date the notice was put online, and the `DT_APPLICATION` field, which represents the time limit for applications. We use these fields to compute the advertising period, expressed in days, which we store in field `publicityDuration` of our database.

Contract Duration Period This period indicates the duration of the contract in months: in the case of a framework agreement, it indicates the length of time during which a contract can be performed. It is directly available as the `DURATION` field in the contract notices. We simply include it unaltered as field `contractDuration` of our database.

Renewals Opportunity Contract notices contain the `RENEWALS` field, which represents the possibility to renew a contract. We simply include it unaltered as the field `renewal` of our database.

3.4 Agent Processing

We apply several distinct processes to agent-related data, in order to populate tables `Agents`, `LotBuyers` and `LotSuppliers`. We describe how we handle location information in Section 3.4.1 and agent names in Section 3.4.2.

As mentioned in Section 2.3.7, the agent SIRET, which should constitute its unique identifier for the French data, is generally not filled in the TED, in practice. For this reason, we define our own identifier. This requires a specific processing aiming at merging occurrences of the same agent appearing under different surface forms, which is described in Section 3.4.3.

3.4.1 Location Information

The goal of the operations described in this section is to solve the *Address Inconsistency* problems identified in Section 2.3.4.

Zipcode and City Normalization As explained in Section 2.3.4, certain fields contain irrelevant information (what we call *Address Pollution*). To solve this issue, we first remove the following information from the city field:

- `CEDEX`, `SP` and `CS`, which is postal information and should not be in this field;
- digits;
- punctuation.

We use regex (regular expressions) to perform this task. Here is an example of the same field before and after this process:

```
ID_NOTICE_CAN   CAE_TOWN
     20113493   MARSEILLE CEDEX 9
     20113493   MARSEILLE
```

During this step, we also partly deal with the *Typographic Inconsistency* problem identified in Section 2.3.4 (inconsistent use of hyphens and diacritics).

Second, we perform a similar task on the zipcode, by removing every non-digit character.

Third and finally, we deal with entries possessing a city name but no zipcode. We leverage a public database called `Hexaposte`[19], which contains the zipcode of each city in France. We use it to retrieve the missing zipcodes. Here is an example:

```
ID_NOTICE_CAN   CAE_TOWN   WIN_POSTAL_CODE
      2010238   PARIS      -
>>    2010238   PARIS      75000
```

Address Normalization Next, we finish dealing with the issues from Section 2.3.4 by normalizing the agents' addresses. First, we remove the different punctuation marks by using a regex, and turn everything to upper case. We also remove all extra spaces. This finishes solving the *Typographic Inconsistency* problem.

Then, we turn to the *Type Confusion* problem (the TED confuses geographic and postal addresses). We remove some words (`CEDEX`, `CS`, `bis`, etc.), especially related to postal addresses, which are not needed or useful in the rest of the process. Here is an example of such a deletion:

```
ID_NOTICE_CAN   CAE_ADDRESS
      2010211   1 PLACE ROBBERT GALLEY BP 9
>>    2010211   1 PLACE ROBBERT GALLEY
```

Finally, in the TED, certain addresses extend over several street numbers, which makes later comparisons more difficult and likely to results in mismatches. Therefore, we use a regex to keep only one street number per address when populating our database. Here is an example of address with multiple street numbers, before and after this processing:

```
ID_NOTICE_CAN   CAE_ADDRESS
      2010869   29-31 COURS DE LA LIBERTE
>>    2010869   29 COURS DE LA LIBERTE
```

At this stage, the *Monolithic Address* problem (constituting elements of the address all forced into the same field) is still open. We solve it later when matching the TED addresses to the SIRENE ones (Section 4.2.3).

3.4.2 Agent Names

The goal of the operations described in this section is to solve the *Joint Agent Description* problem identified in Section 2.3.2 and the *Typographic Inconsistency* problem from Section 2.3.3.

Name Normalization This process concerns the agent names and aims at solving the *Typographic Inconsistency* problem from Section 2.3.3. It involves several steps. First, we remove the different punctuation marks by using a regex and turn everything to upper case. We also remove all extra spaces.

Second, we delete all the information between parentheses, which is generally irrelevant. Here is an example of such a deletion:

[19] https://www.data.gouv.fr/en/datasets/base-officielle-des-codes-postaux/

```
ID_NOTICE_CAN   CAE_NAME
     20102390   AGENCE NATIONALE DES FREQUENCES (ANFR)
>>   20102390   AGENCE NATIONALE DES FREQUENCES
```

Multiple Name Splitting The goal of this process is to solve the *Joint Agent Description* problem from Section 2.3.2, i.e. to separate several agent names involved as buyers and/or winners in the same lot, but expressed as a single string in each concerned field: name, address, zipcode, city, SIRET. Solving this issue requires extracting the appropriate information from each field.

For this purpose, we first leverage the official delimiter, which is the triple hyphen (---). When this delimiter is used in the name field, it also appears in the other fields (address, zipcode, city, and possibly SIRET). It can therefore be used to split each field and retrieve the appropriate information for each concerned agent.

Second, we consider an alternate delimiter: the slash (/). However, we only look for the slash in winner names. Indeed, in CAEs, there are cases where a slash in the name indicates additional information and not a new agent, such as here:

```
ID_NOTICE_CAN   CAE_NAME
    201480448   CEA/Grenoble
```

When the slash is used in the winner's name, the other fields (address, zipcode, city, SIRET) are generally incomplete. Typically, only the first agent is properly described. In this case, the only thing we can do is assign this information to this agent in our database, and leave these fields blank for the other agents, to be filled in the later stages of our process.

We find $37{,}654$ lots (3%) containing at least one of these separators in the data. After splitting all concerned buyer and winner names, the number of agent occurrences passes from $2{,}761{,}930$ to $3{,}017{,}058$, i.e. an increase of $255{,}128$ (9%) occurrences.

3.4.3 Agent Merging

This section aims at sketching how we solve the *Multiple Proper Nouns* and *Name Pollution* problems identified in Section 2.3.3. Our method relies on a temporary table containing multiple forms of the same agents, that we reduce to our final `Agents` table through iterative merging.

Temporary tables In the TED, the same agent is likely to appear under various *forms*. We want to merge distinct forms corresponding to the same agent, in order to get a unique representation of each agent. For this purpose, we use two temporary data tables.

The first is `AgentsTemp`, which initially contains all the data describing the agents.

- `idAgentBase`, unique identifier of each entity entry in the TED.
- `nameAgent`: principal name of the agent.
- `siretAgent`: SIRET of the agent.
- `addressAgent`: address of the agent.
- `cityAgent`: city of the agent.
- `zipcodeAgent`: zipcode of the agent.
- `sameAgent`: identifiers of the other forms of this agent.

It has the same fields as `Agents`, except the primary key, which has a different name (`idAgentBase` instead of `idAgent`), and the additional field `sameAgent`, which connects the various forms of the same agent.

During our merging process, this temporary table is gradually reduced, with fewer and fewer entries in the database, but more complete `sameAgent` fields. At the end of the process, each agent should have a single form in our table, which is then copied in the `Agent` table.

The second temporary table, `AssociationsTemp`, models the association between each lot and the involved buyers and winners in `AgentsTemp`. Its purpose is to later fill the tables

`LotBuyers` and `LotSuppliers` with the appropriate agent identifiers. This table contains the following fields:

- `lotID`: identifier of the lot.
- `idAgentBase`: the identifier of the entity entry in the TED.
- `Type`: The type of the agent, which can be CAE or WIN.

Merging Process During our process, we group agent forms based on their SIRET or cluster. We explain later, in Sections 4 and 5, exactly how these SIRET and clusters are obtained. For now, we describe the generic part of this processing.

When several forms are grouped together, we keep the most probable name, which is the one that is most frequent among these forms. The other names are stored in the `Names` table. For the address, if we find a SIRET, we keep the address found in SIRENE. Otherwise, we apply the previous method for each field.

Here is an example of merging several forms of the same agent:

```
WIN_NAME                          WIN_NATIONALID      WIN_ADDRESS
Eiffage
Eiffage Energie Thermie EST       34002322500055      1 rue Mendes France
Eiffage Energie Thermie Est       34002322500055      1 rue Mendes France
Eiffage Energie Thermie Grand Est 34002322500055      1 rue Mendes France
Eiffage Energies                  34002322500055      1 rue Mendes France
Eiffage Thermie                   34002322500055      1 rue Mendes France
Eiffage Thermie EST               34002322500055      1 rue Mendes France
Eiffage Thermie Est               34002322500055      1 rue Mendes France
Eiffage Thermie Est SAS           34002322500055      1 rue Mendes France
Eiffage energie thermie Grand Est 34002322500055      1 rue Mendes France
Eiffage thermie                   34002322500055      1 rue Mendes France
Eiffage énergie                   34002322500055      1 rue Mendes France

WIN_NAME                          WIN_NATIONALID      WIN_ADDRESS
EIFFAGE THERMIE EST               34002322500055      1 RUE MENDES FRANCE
```

4 Step 2: Agent Identification

As explained in Section 2.3.7, the SIRET is used as a unique identifier to identify economic agents in the French TED dataset. However, this information is missing in most entries. In this step, our objective is to identify as many agents as possible, and fill these missing values. In order to fulfill this task, we need an external data source, since the information of interest is missing from the TED. We use the SIRENE database, which is maintained by the French state. It lists all French companies, and describes them using a variety of fields. We introduce this important tool in Section 4.1.

To retrieve the SIRET from SIRENE, we use the individual information available in the dataset, i.e. the name and the address of the agents. But, as explained is Section 2.3, these fields themselves are not always filled: sometimes there is just the name, sometimes the city is present too, and sometimes the full address, which is composed of the street number, street type, and street name. To solve our issue, we propose several processing steps, that we describe in Section 4.2. Finally, we use a part of our data to assess the performance of our method in Section 4.3.

It is worth stressing that only French agents have a SIRET number. Therefore, this step does not concern agents from foreign countries in the database. These are not considered when assessing the performance of our methods.

4.1 SIRENE Database

The SIRENE database[20] (*Système National d'Identification et du Répertoire des Entreprises et de leurs Etablissements* – National identification system for commercial entities and their facilities) lists all economic agents participating in public procurement, in France. The database was created in 1973[21], but the use of SIRETs became compulsory only in 1997[22]. SIRENE is a large base, containing about 28 million entries. It covers each year since 1973, and includes not only agents that are currently active, but also agents that are no longer active. It is publicly available online since 2017[23].

In this section, we first discuss a specificity of SIRENE: it distinguishes between two levels of economic agents (Section 4.1.1). We then describe the structure of this database (Section 4.1.2). We conclude with a presentation of the processing we applied to its data, in order to make them suitable to our needs (Section 4.1.3).

4.1.1 Entities vs. Facilities

It is important to stress that SIRENE distinguishes two levels of economic agents: entities vs. facilities. *Entities* (or *Unités*, i.e. units, in the SIRENE terminology) are companies, government agencies, department, charity, institutions (legal entity) or people (natural person) that have a legal existence and the ability to enter into agreements or contracts. *Facilities* (or *Établissements* in the SIRENE terminology) are geographically located units where all or part of the entity economic activity is carried out. Agents from the TED correspond to facilities: we want to identify their SIRETs.

Each entity is identified through a unique 9-digit number called the SIREN (*Système d'Identification du Répertoire des Entreprises* – Identification system of the entity register), whereas for a facility it is a 14-digit number called the SIRET (*Système d'Identification du Répertoire des Etablissements* – Identification system of the facility register). The first 9 digits of the SIRET correspond to the SIREN of the associated entity, while the last 5 digits are called

[20] https://www.sirene.fr/
[21] https://www.legifrance.gouv.fr/loda/id/LEGITEXT000006062081/
[22] https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000201066/
[23] https://www.sirene.fr/sirene/public/static/open-data

the NIC (*Numéro Interne de Classement* – Internal classification number) and are specific to each facility. Two facilities linked to the same entity share the same SIREN, but have a different NIC, and therefore a different SIRETs. If an entity closes a facility and reopens it later at the same location, it gets a different NIC.

Here is an example of two facilities related to the same entity:

CAE_NAME	CAE_TOWN	SIRET
ELECTRICITE DE FRANCE	DIJON	55208131788047
ELECTRICITE DE FRANCE	CRETEIL	55208131788054

Before 2005, the prefix of the SIREN number of a *public sector* entity was significant: the first two digits represented its legal category, and the following two characters represented the department code of its head office, for entities with a territorial competence. Here, the notion of *department* refers to a French administrative subdivision, corresponding to the NUTS3 level in the European typology. Table 4 lists these codes and their meaning. The *Legal Category* column refers to a code defined by the INSEE[24], the French institute for statistics, and identifying precisely the type of entity. It is available in the SIRENE database.

SIREN Prefix	Legal Category	Description
10/11	7111–7113	State Administration
12	7120	Central department of a ministry
16	7160	Decentralized department of a ministry
17	7171–7179	Decentralized department of a Region
18	7381–7490	Other public institution
19	7383-7384	Scientific institution or college
21	7210,7312–7314	Municipality
22	7220–7229	Department
23	7230	Region
24	7341–7349	Community of Communes
25	7351–7356	Intercommunal household
26	7361–7366	CCAS and hospitals
27	7371	Public housing office
28	7372–7379	Public administrative establishment

Table 4. Categories of public entities distinguished in SIREN identifiers.

However, since November 2005, only the first two SIREN digits are significant, in two cases:

- Code 13: state administration and agencies with *national* competence.
- Code 20: entities with *territorial* competence.

We have observed that the same municipality can be associated to several different SIRENs, even without changing its address. This happens when several municipalities merge into one: a new SIRENE number is assigned to the newly created larger municipality, whose name is often that of the most important of its constituting smaller municipalities. Here is an example of such situation:

NAME	ADDRESS	TOWN	SIRET
COMMUNE DE THORAS	MAIRIE	THORAS	20005932700060
COMMUNE DE THORAS	MAIRIE	THORAS	21430245700012

The municipalities of `CROISSANCES` and `THORAS` merged on the 1st of January 2016, which resulted in a new and larger municipality, also called `THORAS`. It has exactly the same name, address, and information than the older and smaller `THORAS`, but possesses a different SIRET.

[24] https://www.insee.fr/en/accueil

4.1.2 Structure of the Database

SIRENE is accessible via 3 methods: first, a dedicated website allows a human access; second, it can be accessed programmatically through an online API; and third, it is possible to download the database as CSV files, in order to use them locally. Like for the TED, we adopt the last method, because it allows us to have more control over the way economic agents are searched in these data.

The CSV version of the database consists of four parts:

- `StockUniteLegale_utf8.csv`: a CSV file containing all the entities (unités) (22M entries), be them open or closed, with the latest available information, including the SIREN, usual denomination and acronym.
- `StockEtablissement_utf8.csv`: a CSV file containing all facilities (établissements) (28M entries), be them open or closed, with the latest available information, including the SIRET, zipcode, city, address and trading name. It should be noted that in SIRENE, an address is represented by 3 fields:
  - `typeVoieEtablissement`: street number.
  - `numeroVoieEtablissement`: type of road.
  - `libelleVoieEtablissement`: street name.
- `StockEtablissementHistorique_utf8.csv`: a CSV file containing the historical modifications of the facilities, including their opening and closing dates.
- `StockUniteHistorique_utf8.csv`: a CSV file containing the historical modifications of the entities, with the previous names of each entity.

Using these four files, we create a temporary database containing four tables, as described in Figure 7. Its goal is only to cross-reference the agents from the TED with the facilities from SIRENE, in order to fill the missing SIRETs. It is not meant to stay in our database after the completion of this task.

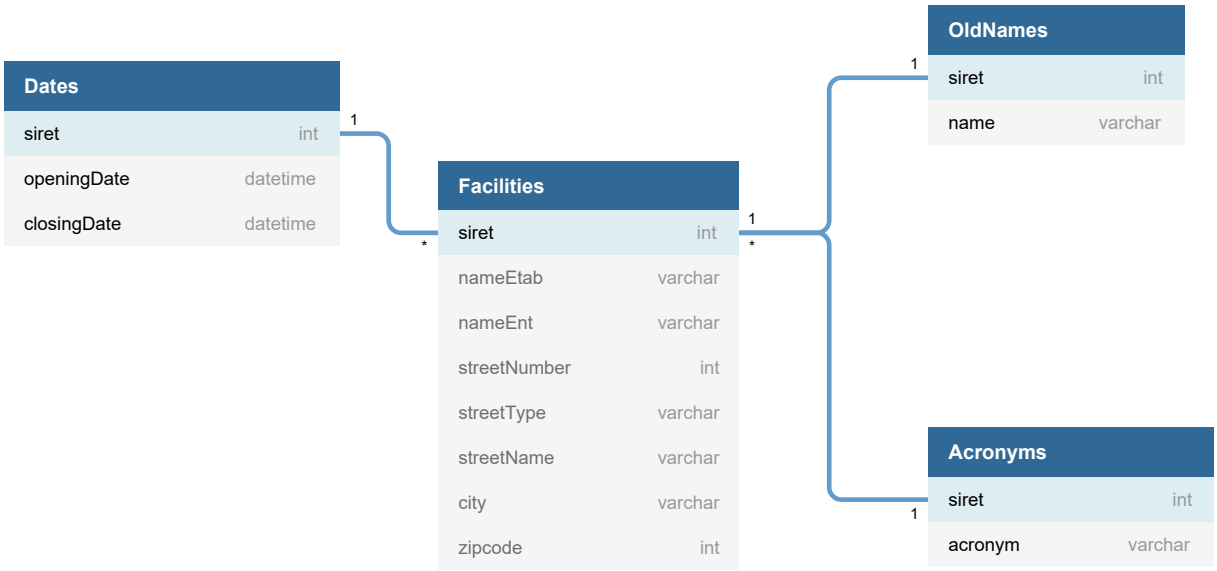


Figure 7. Structure of the SIRENE database, shown as an Entity-Relation diagram.

Table `Facilities` It contains the facilities (établissements), described using the following fields:

- `siret`: SIRET, i.e. unique identifier of the facility.
- `nameEtab`: name of the facility.
- `nameEnt`: name of the associated entity.
- `streetNumber`: street number.

- `streetType`: type of the street.
- `streetName`: name of the street.
- `town`: city.
- `zipcode`: zipcode.

Table `Dates` One given facility can open and close several times during its existence. For this reason, it is not possible to store opening and closing dates directly in the Facilities table: we use the `Dates` table for this purpose. It contains the following fields:

- `siret`: SIRET of a facility.
- `openingDate`: opening date of this facility.
- `closingDate`: closing date of this facility.

The closing date is missing when the facility is still open.

Table `OldNames` An entity can change its name during its existence. File `StockUniteLegale_utf8.csv` only provides the latest name, whereas file `StockUniteHistorique_utf8.csv` contains the older ones.

We proceed similarly in our database: table `Facilities` contains the latest name (field `nameEnt`), whereas the previous ones are stored in table `OldNames`. The latter contains the following fields:

- `siret` SIRET of a facility.
- `name`: one of the previous names of this facility.

Table `Acronyms` Some entity names take the form of acronyms. In order to leverage them later when comparing agent names, we gather all these specific names in a different table called `Acronyms`, and containing the following fields:

- `siret`: SIRET of a facility.
- `acronym`: acronym of this facility.

4.1.3 Preparation of the Data

Overall, the SIRENE data appear to be of good quality, and does not require much preparation. The only issues that we detected concerns the names of its entities and facilities. Indeed, as mentioned before, we want to cross-reference the unidentified agents from the TED with the facilities present in SIRENE based on their names and addresses. However, the facility names are not always filled in SIRENE, and when they are, they are not always the most appropriate field for our task, as SIRENE may contain several names. Moreover, SIRENE tend to contain full names, when the TED sometimes contains acronyms instead.

In SIRENE For each entity in SIRENE, one name is possibly stored in the following fields:

- `denominationUniteLegale`: name in case of legal person.
- `denominationUsuelleUniteLegale`: name commonly used by the public.
- `nomUniteLegale`: name in case of a natural person.
- `prenomUniteLegale`: first name in case of a natural person.
- `sigleUniteLegale`: acronym of the facility name.

For each facility in SIRENE, we have a single name stored in the following field:

- `enseigneEtablissement`: name of the facility.

Moreover, as mentioned before, an entity may change its name over time. For instance:

```
    SIREN   UNITE NAME
247400161   SIVOM MORILLON SAMOENS SIXT VERCHAIX
247400161   SIVOM EAU ASSAINISSEMENT MOR/SAM/VER/SIX
247400161   SI DES MONTAGNES DU GRIFFE
```

To create the table `Facility`, we link the facilities in SIRENE to their entities, in order to retain for each facility both the company name and the facility name. We extract opening dates, acronyms and older names of the historical CSV to create the 3 other tables.

In FOPPA Some agents appear under their full name in SIRENE, whereas their acronym is used in the TED. This is particularly the case for education and medical facilities, for instance:

ID_NOTICE_CAN	CAE_NAME	Name in SIRENE
2010332	CH de Belfort Montbéliard	Centre Hospitalier de Belfort Montbéliard

Table 5 shows the number of agent occurrences named after common hospital acronyms (`CH`, `CHD`, `CHU`, `CHR`) in the TED. Almost half of them are described using their full name in SIRENE.

Agent Occurrences	Centre Hospitalier	CH	CHD	CHU	CHR	Total
Count	124,484	32,313	2,049	59,428	3,282	221,556
Proportion	56.19%	14.58%	0.92%	26.82%	1.48%	100.00%

Table 5. Different expressions referring to hospitals, and their respective frequencies.

We handle this issue by replacing these common acronyms by the corresponding full string in our database, in order to ease name comparison during the siretization process.

4.2 Matching Algorithm

In this section, we describe our proposed method to match an unidentified economic agent from the TED to a facility from SIRENE, and therefore obtain the agent's SIRET. As mentioned before, this task is part of what we call *siretization*, i.e. the retrieval of missing SIRETs.

Our algorithm is described in Figure 8. Each green block represents a subset of facilities from the SIRENE database. The first subset is initialized by selecting facilities which are compatible with the information provided by the considered lot description from the TED dataset. We reduce this set of potential candidates by filtering them depending on the other available fields, through three phases:

1. *Date & Domain filtering phase*: we use SQL queries in order to find valid candidate facilities in SIRENE, leveraging the TED agent's activity domain, opening dates and department.
2. *Name filtering phase*: we perform an approximate matching based on the names of these valid facilities in order to refine the set of potential candidates.
3. *Location filtering phase*: we perform an approximate matching based on the address, city and zipcode of each potential candidate, in order to find the most likely ones.

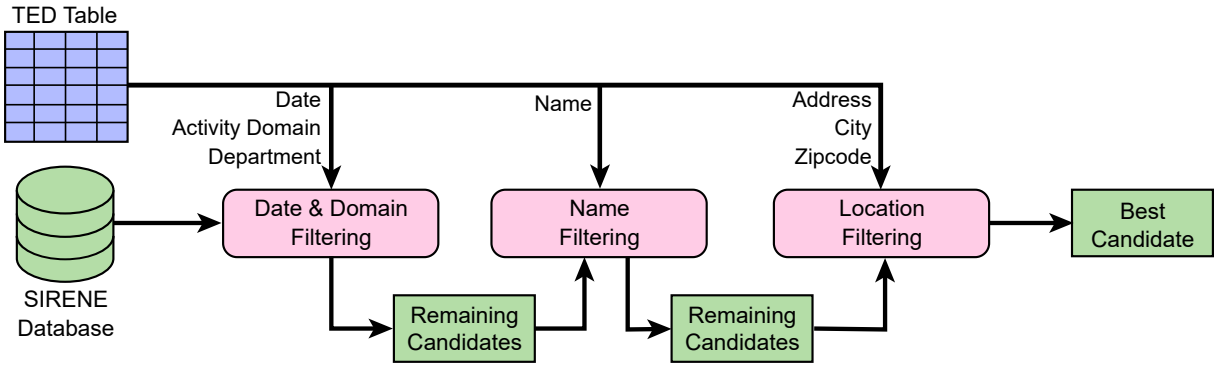


Figure 8. Successive filtering phases of the agent identification process. The square boxes represent data, and the round ones processing steps.

In then end, the process outputs the single SIRENE candidate best matching the TED agent. It is possible that the process stops before that point, if no suitable candidate is found during one of the filtering phases.

We could directly try to match the TED agent's name to the whole SIRENE database, but this has several drawbacks. First, this is computationally expensive, as SIRENE contains millions of entries. Second, this would lead to numerous errors, as many agents are likely to have similar names while differing on other characteristics (especially their location). For this reason, the first phase (Section 4.2.1) aims at reducing the number of candidates before performing the name matching in the second phase (Section 4.2.2), in order to select good candidates. Finally, the third phase (Section 4.2.3) aims to rank these candidates in order to select the best one.

4.2.1 Date & Domain Filtering

Each filtering phase focuses on a specific type of information describing the agent in the TED. The first uses temporal and activity-related information, as well as a part of the geographical information and possibly certain aspects of the name:

- Department (first 2 digits of the zipcode);
- Activity domain;
- Opening dates;
- Name.

We first filter by department, using only the first two digits of the zipcode field (`POSTAL_CODE`). Based on our observation of the TED data, it is the most reliable geographical information, and it allows us to greatly reduce the number of candidates. Similarly, we only retain the SIRENE facilities which are related to the activity domain of the targeted TED agent, and which are active at the date of the considered contract.

In addition, we use human knowledge to identify situations allowing to narrow the candidate set even further. The names of certain facilities contain predefined terms that characterize their general nature or activity domain. For example, there are only a few ways to refer to a hospital in France, depending on its role and importance:

Common term	Acronym associated
Centre Hospitalier Regional	CHR
Centre Hospitalier Departemental	CHD
Centre Hospitalier Universitaire	CHU
Centre Hospitalier General	CHG

We leverage these terms to constrain the set of candidates even further. In the previous example, this means only searching among the hospitals contained in SIRENE. This allows to significantly decrease the number of candidates.

4.2.2 Name Filtering

The next phase exclusively focuses on the TED agents and SIRENE facility names. We refine the potential SIRENE candidate set obtained at the previous phase, by retaining only the remaining facilities whose name is close enough to the TED agent's name.

For this purpose, we perform an approximate comparison between the TED and SIRENE names, through a method based on the Levenshtein distance [15]. We use the Python library `Fuzzywuzzy`[25], which proposes 4 main string comparison methods:

- `ratio`: simple Levenshtein distance, which is normalized by dividing by the length of the string.
- `partial_ratio`: comparison between the shortest name and all the substrings of the same length found in the longer name.

[25] https://github.com/seatgeek/thefuzz

- `token.sort_ratio`: both names are tokenized, these tokens are sorted alphabetically, then concatenated, before computing the Levenshtein distance on both resulting strings.
- token.set_ratio: same operation as `sort_ratio`, but the common tokens are taken out.

Each of these functions returns a score between 0 (completely different) and 100 (perfectly identical). In the rest of this document, we call this score *similarity*.

As mentioned in the Section 4.1.2, we have at most 4 possible types of names to characterize a SIRENE facility: facility name, entity name, previous entity names, acronyms. We use different similarity functions depending on the types of the available names.

Acronyms We distinguish between two cases: either the TED name field only contains an acronym, or it contains an acronym and some additional text (full name, administrative subdivision, service, department, etc.). Based on empirical estimation, we assume that names shorter than 7 characters generally correspond to the first case (sole acronyms). We handle both cases differently, similarity-wise:

- Sole acronym: we use the `ratio` function, and keep only results with a maximal score, i.e. 100.
- Acronym and other text: we use the `partial_ratio` function, and keep only results with a maximal score.

A slight difference between two acronyms is absolutely not a guarantee of proximity between two companies. Here is an example showing two similar acronyms referring to completely different companies:

SIRET	ACRONYM
48760448000029	EDG
55208131766522	EDF

This is the reason why, in this case, we perform *exact* comparison by retaining only maximal similarity cases.

Other names We use the same function `token_set_ratio` for all other types of names. The only difference lies in the threshold that we set for keeping candidates or not:

- For the cases where we used human knowledge (hospitals, department etc.) at the previous stage (Section 4.2.1), the threshold must be high, since each of the remaining candidates is likely to have a similar name or at least some words in common. We chose an acceptance threshold of 90, which gives suitable results according to our experiments.
- For other cases, we found that the best acceptance threshold is around 70.

Overall Result Based on various approximate comparisons, our approach handles the *Name Pollution* problem (agent names containing irrelevant information) identified in Section 2.3.3. When a facility has several names in SIRENE, we treat each one separately using the above methods. We then keep the highest score as the result of the comparison.

4.2.3 Location Filtering

The last filtering phase takes advantage of the rest of the geographical information, in order to filter the candidate facilities remaining after the previous phase (name filtering):

- City;
- Complete address, i.e. street number, street type and street name.

There are two situations that complicate the comparison of the TED and SIRENE addresses. First, in TED, general address information sometimes appears in the `city` field, an issue that we call *Address Pollution* in Section 2.3.4. Here is an example:

ID_NOTICE_CAN	CAE_TOWN
2016156574	LA DEFENSE

In this case, `LA DEFENSE` is not a town but a business district in Paris. Consequently, only matching the city will not return any result.

Second, TED does not always provide all 3 pieces of address information: the number may be missing, or the type of street may be different, etc. However, a single error on one of these three fields does not necessarily invalidate the whole address.

Therefore, in order to perform this comparison, we concatenate all the fields in one string. This means that on the TED side, we merge `address` and `city`; and on the SIRENE side, we merge `streetnumber`, `streettype`, `streetname`, and `city`. This allows making the most of the available information. In the previous example, our method is able to factor the district in the comparison, as it appears in the SIRENE `address` field. Our approach also allows taking care of the *Type Confusion* (mixing geographic and postal addresses) and *Monolithic Address* (combinig various parts of the address in the same field) issues identified in Section 2.3.4.

Based on the concatenated string, we compute a score for each remaining candidate using function `token_set_ratio`. We then take the average between this score and the one obtained at the previous phase (name filtering). Our goal is to boost candidates whose names are very similar to the target's. The candidate with the highest average score is selected as the final result.

4.3 Performance Assessment

In this section, we assess the performance of our identification method by comparing its output with two distinct ground truths. On the one hand, we leverage TED entries whose SIRET is filled out in the TED (Section 4.3.1). However, we observed that the agents concerned by these cases are mainly buyers, which suggests a potential bias, agent-wise. This is why, on the other hand, we consider a random sample of entries without SIRET, which we identify manually (Section 4.3.2). Finally, we extrapolate these performance scores to the whole dataset, and discuss briefly the results of this step (Section 4.3.3).

4.3.1 Known SIRETs

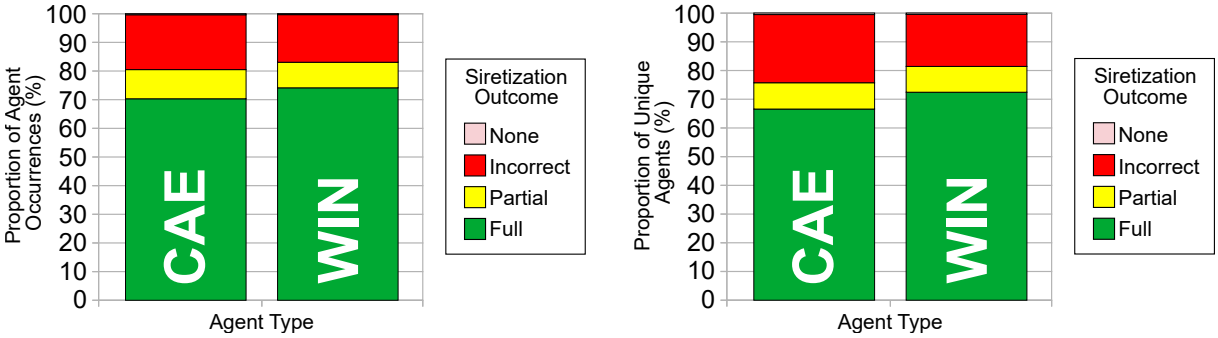
If we count the number of agent occurrences (by opposition to unique agents) with a valid SIRET, we get $207{,}316$ buyers and $40{,}034$ winners. We remove these SIRETs from the database and apply our method, to check whether it can recover them.

In order to assess the performance of our method, we consider 4 different outcomes:

- *Full SIRET*: the method correctly retrieves the SIRET, i.e. all 14 digits. The matching is then successful.
- *Partial SIRET*: the method only retrieves the SIREN part of the code, i.e. the first 9 digits. Put differently, the SIRENE entity is correct, but the method fails to identify the facility.
- *Incorrect SIRET*: the method selects an incorrect candidate and returns a completely incorrect SIRET.
- *No SIRET*: the method fails to identify any suitable candidate, and returns no SIRET at all.

Figure 9 presents the results of our evaluation, for the SIRETs that originally exist in the dataset. The x-axis represents the agents type: the left-hand bar focuses on the buyers and the right-hand one on the winners. The colors represent the four outcomes described before: *Full SIRET* (green), *Partial SIRET* (yellow), *Incorrect SIRET* (red), and *No SIRET* (pink). The y-axis represents the percentage of agent occurrences for each outcome. The exact values are listed in Appendix E.1: Table 44 for Figure 9a and Table 45 for Figure 9b.

For these agents, whose SIRET is already known from the TED, our method reaches complete success, i.e. identification of the full SIRET, for $70{,}29\%$ of the buyer occurrences and 74.11% of the winner occurrences. For buyers, we consider that a partial SIRET is also a good result, since procurement management is often centralized at the main entity. Summing up



(a) Results for agent occurrences.

(b) Results for unique agents.

Figure 9. Distribution of the four possible outcomes of the agent identification process for the *predefined* SIRETs.

the two, we get 80.47% of success. The result on the winners is higher, with 83.03%. This is explained by the fact that the awards notices are always filled by the buyers.

When they bother filling their own information, they tend to get their SIRET right, but not pay too much attention to the rest of their fields. Alternatively, certain large agents such as ministries tend to provide some unnecessary information in the name of address fields, such as directions, making their identification more difficult. On the contrary, when the buyers fill the winner's SIRET, they generally tend to make sure that they get its other fields right, too.

Among the $207{,}316$ buyer occurrences, the TED contains only $6{,}115$ unique buyers. For the winners, we have $40{,}034$ occurrences vs. $18{,}209$ unique values. Figure 9b represents the result of our evaluation for unique agents. The performance is lower than when considering occurrences: 66.56% for buyers and 72.42% for winners in terms of full SIRETs; 75.72% for buyers and 81.41% for winners if we add partial SIRETs. This is due to the greater prevalence of small agents, i.e. agents that rarely appear in the data. These agents tend to exhibit more problems in terms of both completion and reliability of the provided information. They are consequently more difficult to identify.

4.3.2 Manually Annotated SIRETs

To constitute the second ground truth, we first randomly sample 500 agents (250 buyers and 250 winners) from the unidentified entries of the TED that possess both a city and a name. This sample does not contain multiple occurrences of the same agent, because of its small size. Second, we take advantage of these two fields (name and city) to *manually* retrieve the missing SIRETs of all 500 agents.

This method allows solving the bias present in the predefined SIRET dataset (Section 4.3.1), i.e. the overrepresentation of CAEs. The evaluation method is the same as in Section 4.3.1. Figure 10 presents the obtained results, whereas the exact values are shown in Table 46 (Appendix E.1).

The proportion of fully identified SIRETs is 71.2% for buyers and 62.8% for winners. If we add partially identified SIRETs, the proportion reaches 80.8% and 73.2%, respectively. Compared to the results obtained for the predefined SIRETs, these performances are better than for unique agents (Figure 9b) and worse than for agent occurrences (Figure 9a). As we randomly select agent occurrences from the TED to constitute the annotated dataset, there is a higher chance to get large agents, since they are more frequent in the database: this could explain this observation.

There is a larger number of agents for which the algorithm is not able to return a SIRET. This is because these agents originally have many missing fields in the TED and/or a poorly written name (compared to the one present in SIRENE). Unlike before, we obtain better

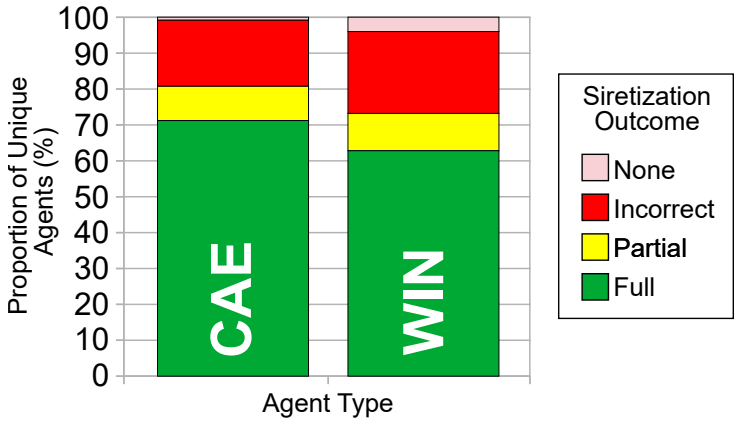


Figure 10. Distribution of the four possible outcomes of the agent identification process for the *annotated* SIRETs (the exact values are shown in Table 46).

performances for buyers than for winners: this is because this time, the ground truth is based on agents without SIRET. Buyers that do not fill their winner's SIRET tend to provide incorrect information regarding the rest of the winner's information, or no information at all, making the identification task more difficult.

4.3.3 Extrapolation

Assuming that the performance scores obtained at Sections 4.3.1 and 4.3.2 apply to the whole dataset, we can extrapolate to estimate the overall performance of our identification step, in terms of both agent occurrences and unique agents.

Agent Occurrences After all the processing described in Section 3, the dataset contains a total of 2,878,683 agent occurrences. Among them, only 649,170 possess a structurally valid SIRET. This means that the 2,229,513 remaining occurrences either have a structurally invalid SIRET or no SIRET at all. The application of the identification step to these occurrences allows us to retrieve 2,025,745 new SIRETs, including 995,745 buyers and 1,030,000 winners. Table 6 summarizes the situation.

Agent Type	Known	Siretized	Not Siretized	Total
Buyer occurrences	477,421	995,745	24,248	1,497,414
	31.88%	66.50%	1.62%	100.00%
Winner occurrences	171,749	1,030,000	179,520	1,381,269
	12.43%	74.57%	13.00%	100.00%
Agent occurrences	649,170	2,025,745	203,768	2,878,683
	22.55%	70.37%	7.08%	100.00%

Table 6. Status of the SIRET of *agent occurrences* after the siretization: SIRET already known before this step (Column *Known*), SIRETs estimated during the identification step (*Siretized*), and SIRETs that could not be estimated during this step (*Not Siretized*).

We now focus only on the agent occurrences identified at this step, and estimate the proportions of correct and incorrect SIRETs based on the performance scores presented in Sections 4.3.1 and 4.3.2. For this purpose, we use the proportions of fully, partially, and incorrectly retrieved SIRETs. We ignore the proportion of SIRETs that could not be retrieved, and consequently normalize the three other proportions so that they sum to one. According to our calculations, these proportions are respectively 70.57%, 10.22%, 19,21% for buyers, and 74.32%, 8.96%, 16.72% for winners. Table 7 shows the results of this extrapolation.

We consider both fully and partially recovered SIRETs as correct. If we combine these

Agent Type	Full	Partial	Incorrect	Total
Buyers occurrences	702,656	101,764	191,325	995,745
	70.57%	10.22%	19.21%	100.00%
Winner occurrences	765,506	92,323	172,171	1,030,000
	74.32%	8.96%	16.72%	100.00%
Agent occurrences	1,468,161	194,087	363,497	2,025,745
	72.48%	9.58%	17.94%	100.00%

Table 7. Extrapolation of the identification performance to the whole set of *agent occurrences* that could be siretized during this step.

estimations with the SIRETs that were originally known in the TED data, we get the scores shown in Table 8. Our estimation shows that 80.29% of the agent occurrences have a correct SIRET, 12.63% have an incorrect SIRET, and 7.08% have no SIRET at all.

Agent Type	Correct	Incorrect	Missing	Total
Buyers occurrences	1,281,841	191,325	24,248	1,497,414
	85.60%	12.78%	1.62%	100.00%
Winner occurrences	1,029,578	172,171	179,520	1,381,269
	74.54%	12.46%	13.00%	100.00%
Agent occurrences	2,311,418	363,497	203,768	2,878,683
	80.29%	12.63%	7.08%	100.00%

Table 8. Estimated completion of the SIRET field in *agent occurrences* after the identification step, for the whole dataset.

Unique Agents By comparison, before the identification step, we have 98,574 unique buyers and 633,790 unique winners, corresponding to a total of 732,364 unique agents. Among them, 41,935 possess a structurally valid SIRET, whereas 690,429 do not, and have to go through the identification process. It is able to retrieve a SIRET for 614,435 of them, leaving the 75,994 remaining unique agents without a SIRET. Table 9 summarizes the distribution of buyers vs. winners.

Agent Type	Known	Siretized	Not Siretized	Total
Unique buyers	14,592	81,704	2,278	98,574
	14.80%	82.89%	2.31%	100.00%
Unique winners	27,343	532,731	73,716	633,790
	4.31%	84.05%	11.63%	100.00%
Unique agents	41,935	614,435	75,994	732,364
	5.73%	83.90%	10.38%	100.00%

Table 9. Status of the SIRET of *unique agents* after the identification: SIRET already known before this step (Column *Known*), SIRETs estimated during the identification step (*Siretized*), and SIRETs that could not be estimated during this step (*Not Siretized*).

It is important to stress that, unlike the total number of agent occurrences, which is constant, the number of unique agents changes at each step of our process, due to the unification of occurrences differing in form while representing the same entity. Consequently, there are two conflicting ways of counting the number of unique agents that could be identified at this step: either by considering the increase of unique agents with a SIRET, or the decrease of those without a SIRET. On the one hand, as mentioned before, agents whose SIRET could be retrieved may get merged, which also affects the count. On the other hand, this is not the case for agents whose SIRET could not be retrieved. For this reason, we adopt

the second method (decrease in unidentified agents) in Table 9.

Agent Type	Full	Partial	Incorrect	Total
Unique buyers	54,831	7,538	19,335	81,704
	67.11%	9.23%	23.67%	100.00%
Unique winners	387,187	48,282	97,261	532,731
	72.68%	9.06%	18.26%	100.00%
Unique agents	442,018	55,820	116,596	614,435
	71.94%	9.08%	18.98%	100.00%

Table 10. Extrapolation of the identification performance to the whole set of *unique agents* that could be identified during this step.

Like for occurrences, we now focus only on the unique agents that could be identified during this step, and estimate the proportions of correct and incorrect SIRETs based on the performance scores presented in Sections 4.3.1 and 4.3.2. The results of this extrapolation are shown in Table 10.

Agent Type	Correct	Incorrect	Missing	Total
Unique buyers	76,961	19,335	2,278	98,574
	78.07%	19.62%	2.31%	100.00%
Unique winners	462,813	97,261	73,716	633,790
	73.02%	15.35%	11.63%	100.00%
Unique agents	539,774	116,596	75,994	732,364
	73.70%	15.92%	10.38%	100.00%

Table 11. Estimated completion of the SIRET field in *unique agents* after the identification step, for the whole dataset.

We then combine these estimations with the SIRETs that were originally known in the TED data, considering both fully and partially recovered SIRETs as correct, like we did for occurrences. Table 11 exhibits our results, showing that 73.70% of the unique agents have a correct SIRET, 15.92% have an incorrect SIRET, and 10.38% have no SIRET at all. These scores are below those obtained when considering agent occurrences instead of unique agents: this shows that frequent agents are better described in the dataset

Depending on how we count, approximately 20% of the agents either have an incorrect SIRET, or not SIRET at all. The objective of the next step, agent clustering, is to handle these remaining issues.

5 Step 3: Clustering-Based Merging

As explained in Section 4.3, our identification process fails to retrieve a reliable SIRET for certain agent occurrences. We assume that a part of these occurrences actually represent the same entities as other correctly identified occurrences, or even other unidentified occurrences. The goal of this step is to group these occurrences under the same entries in our database.

For this purpose, we leverage the Dedupe library, which we describe in Section 5.1. We use this library to perform a cluster analysis of the TED agents, as explained in Section 5.2. We then have to process the clusters produced by Dedupe in order to decide which agents to merge, as described in Section 5.3. Finally, we use a part of our data to assess the performance of this step in Section 5.4.

5.1 Description of Dedupe

Dedupe[26] is a Python library which performs fuzzy matching, record deduplication and entity resolution. Its algorithm is based on three main steps: compute the *record similarity*, use *blocking* to handle large datasets, and perform *cluster analysis* to uncover groups of similar records. Dedupe uses active learning to estimate the best parameter values during all of these steps.

Record similarity In order to compare two strings, Dedupe uses the *Affine Gap* distance [21], a variation of the *Hamming distance* [11]. The Hamming distance is the number of different letters located at the same position in both strings. The affine gap distance allows using gaps between letters (with a penalty), which provides a more flexible matching.

Dedupe compares two records field-by-field, i.e. by considering each field separately, before combining the resulting distance values. It assigns a weight to each field, corresponding to different levels of importance during this comparison. These fields also allow normalizing the overall score in order to get a probability value. These weights are data-dependent, and learned during the active learning phase.

Blocking After assigning the weights, one could theoretically compute the distance between each pair of records. However, it is not possible to do so in practice, as it would be computationally too costly. To solve this problem, Dedupe uses a system called *blocking*: the data are divided in groups of records sharing some common patterns. A pattern can be, for instance, having the same value for a specific field, or the same first characters. Each record can be located in one or more so-called blocks.

A blocking rule focuses on a specific subset of fields, on which it defines a set of constraints (strict equality, but also more flexible comparisons). Each rule defines one block, and records respecting several rules at once belong to the different corresponding blocks.

Once the blocks are created, Dedupe only compares records within the same block, in order to avoid comparisons between records that are too different. The rules for creating blocks are data-dependent, and Dedupe learns them during the training phase.

Clustering The last step consists in forming clusters containing similar records. This task is complicated by the fact that Dedupe does not have access to the similarity of certain pairs of records, if they do not belong to the same block. In order to solve this issue, Dedupe uses a *hierarchical clustering with centroid linkage* [16]. The resulting clusters contain groups of records considered as duplicates. In order to perform the clustering, Dedupe leverages a user-defined *cophenetic threshold* [20], i.e. the minimal similarity value for two records to be placed in the same cluster.

Active Learning Active learning requires the user to provide the tool with annotations on specific cases identified as relevant. These are modeled as a set of pairs that:

[26] https://github.com/Dedupeio/Dedupe

- are duplicates for Dedupe, but not in the same blocking group;
- are not duplicates for Dedupe, but in the same blocking group.

Dedupe provides a pair of this set to the user, who indicate if they are the same agent or two different agents. Thanks to this new labeled example, Dedupe updates the blocking rules, the weights of the algorithm and the set of pairs. Dedupe proposes new pairs to the user, until he or she decides to stop the process.

5.2 Application to our data

After the identification phase, we can distinguish two types of agents:

- Identified agents: each unique SIRET is associated to a single surface form, thanks to the merging step described in Section 3.4.3.
- Unidentified agents: these can be one of the following three cases:
  - Another surface form of an already identified agent that our process did recognize correctly;
  - A surface form of an agent that appears under other unidentified forms;
  - An agent different from all other agents present in the database (identified or not).

We compare each agent using the non-SIRET fields in our database, i.e.:

- `name`.
- `address`.
- `city`.
- `zipcode`.

Active Learning Phase To start, we perform the active learning phase on 500 pairs. We manually identify the pairs selected by Dedupe, which correspond to 78 positive pairs (different forms of the same agent) and 422 negative pairs (not the same agent).

Here are some examples of the blocking rules used by Dedupe:

- Same first 5 characters on the name field.
- Phonetic matching on the address field.
- Same integer on the address field.
- Same six-gram on the city field.

Clustering Phase The next step consists in performing the cluster analysis. We select a conservative cophenetic threshold, because some entries with name and city could be associated despite a difference of city, and thus necessarily of agents. A threshold of 0.8 gives suitable results according to our experiments. After this processing, Dedupe outputs a CSV with 2 additional fields to each agent: a cluster number and a confidence score. The latter is a measure of similarity of the agent of interest, in relation to the other agents in the same cluster.

5.3 Postprocessing

Once we have the Dedupe clusters, we must process them in order to decide which agents must be merged in our database. In the following, we consider all possible situations and the corresponding actions.

Singleton Cluster A singleton contains only one agent. Dedupe did not find any other agent sufficiently similar to put them together.

If the agent has a SIRET, then we assume that there is no other form of the same agent possessing a different SIRET, and that there is no unidentified agent matching it. This SIRET may be incorrect, but at this step, we assume that it is correct. It may be revised at the next step, in case of merging with another identified agent.

If the agent constituting the singleton is unidentified, then the case is a failure, as our process could not identify its SIRET. The next step of our pipeline may succeed later in this

task. The agent is assigned a unique identifier, internal to our FOPPA.

Multiple SIRETs It is possible that Dedupe puts several identified agents in the same cluster. In this case, we have what we call a SIRET *conflict*, since all these agents correspond to the same entity according to Dedupe, yet they are considered as distinct ones according to our database.

We could either consider that our identification was incorrect or that Dedupe improperly considers two distinct agents as duplicates. We choose to favor the former assumption, because our previous experiments show that our identification process can sometimes produce incorrect SIRETs. Typically, when two occurrences of the same agent are poorly filled in the TED, with small disparities between them, the identification process does not lead to the same result when matching with SIRENE. In addition, as mentioned before, we use a conservative cophenetic threshold with Dedupe.

Consequently, we merge the concerned unique agents, using the same strategy as in Section 3.4.3: we keep the most frequent value for each field, including the SIRET. In the case of the SIRET, we also consider the number of occurrences of each agent in the original dataset to determine which unique agent is majority in the cluster. The rationale behind this strategy is to favor more frequent agents, as their information tend to be more reliable in the TED. If the cluster contains unidentified agents, they are also merged during the process.

Other Cases If the cluster contains several agents without any SIRET, then we assume that they are all different occurrences of the same agent. We combine them all, to get a single entry in our database, identified by its own unique internal identifier. If one of the agents has a SIRET, then we also use it to identify the combined entry.

5.4 Performance Assessment

In this section, we assess the performance of the clustering step. We first present some general statistics regarding the size of the clusters identified by Dedupe (Section 5.4.1). Then, we propose two methods to assess the amount of false positives (Section 5.4.2) and false negatives (Section 5.4.3), respectively. The *false positives* are agents placed in the same cluster by Dedupe, when they actually correspond to several distinct entities, and should therefore be located in different clusters. On the contrary, the *false negatives* are agents located in different clusters, when they actually correspond to the same entity, and should therefore belong to the same cluster.

5.4.1 Cluster Sizes

The clustering process distributes the $306{,}984$ unique agents remaining after the identification step over $301{,}096$ clusters. These are small, with an average size of 1.08 agent by cluster. Table 12 shows the full distribution of the cluster sizes, and it appears that most clusters are singletons (94%).

Cluster size	1	2	3	4	5	6+	Total
Count	296,296	4,158	438	118	41	45	301,096
Proportion	98.40%	1,38%	0.14%	0.06%	0.01%	0.01%	100%

Table 12. Distribution of the number of agent occurrences by cluster.

As explained in Section 5.3, singleton clusters do not require any additional work during the post-processing: if they possess a SIRET, then it is assumed correct (for now), and if they do not, it means that Dedupe could not find one. The remaining $4{,}800$ clusters (6%) contain several agents, possibly corresponding to a single or several identifiers (SIRET or SIREN).

5.4.2 False Positives

False positives correspond to SIRET or SIREN conflicts, as defined in Section 5.3. We first present and discuss some statistics related to false positives, computed on the whole dataset. Then, we assess the performance of our clustering step using the ground truth from Section 4.3.1, i.e. the one based on pre-existing SIRETs. Finally, we discuss how these cluster-based results translate in terms of identifier estimation.

Whole Dataset Table 13 represents the distribution of clusters according to their numbers of distinct identifiers (SIRETs and SIRENs). By comparison, Table 12 focuses on agents, not identifiers: some agents have no identifier, and the same cluster can contain several agents sharing the same SIREN.

Number of distinct IDs	0	1	2	3	4	5+	Total
SIRETs	73,751	224,261	2,765	224	51	44	301,096
	24.49%	74.48%	0.92%	0.07%	0.02%	0.01%	100.00%
SIRENs	73,751	225,669	1,634	40	2	0	301,096
	24.49%	74.95%	0.54%	0.01%	0.00%	0.00%	100.00%

Table 13. Distribution of the number of distinct identifiers by cluster, in terms of SIRET and SIREN.

There are $298{,}012$ (99.0%) and $299{,}420$ (99.4%) clusters with no conflict (i.e. they contain 0 or 1 identifier), in terms of SIRETs and SIRENs, respectively. That leaves us with a total of $3{,}084$ and $1{,}676$ conflicted clusters. These numbers respectively amount to 64% and 35% of the $4{,}800$ clusters containing several agents (cf. Section 5.4.1). Among them, $1{,}408$ clusters are conflicted according to SIRETs, but not when focusing only on SIRENs.

There can be two reasons for these conflicts: either the identification process is incorrect and the concerned agents should have the same SIRET, or the clustering step is incorrect and those are indeed different agents that should be kept separated. The reliability of the identification step is already assessed in Section 4.3: here, we want to focus on the latter case.

Known Identifiers In order to investigate the performance of the clustering step, we compute the same statistics as in Table 13, but while focusing only on known SIRETs and SIRENs, as we did for the identification step in Section 4.3.1. Put differently, we use only the identifiers that were originally provided by the TED, thus excluding those resulting from our identification step. There are $24{,}324$ known SIRETs and $21{,}370$ known SIRENs. Our assumption here is that the SIRETs and SIRENs originating from the TED are certainly correct, and should therefore be placed in distinct clusters. As our identification step involves merging all agent occurrences possessing the same SIRET, each known SIRET appears once and only once at the clustering step.

Table 14 shows the distribution of clusters according to the number of known identifiers they contain. We get a total of $22{,}548$ clusters containing at least one such SIRET. According to the table, 96.64% of the known SIRETs are correctly placed in singleton clusters, whereas the rest are incorrectly mixed with other known SIRETs.

When characterizing clusters in terms of their numbers of distinct SIRENs instead of SIRETs, the performance is slightly higher: 98.27%. Put differently, 368 of the 588 known SIRETs confused by Dedupe (i.e. 63%) correspond to facilities belonging to the same entity (according to the SIRENE terminology), i.e. they share the same SIREN. After a manual verification, we conclude that these cases are most likely due to some entities having several facilities located at the exact same place. The following example illustrates this situation for two agents sharing the same name and address, but possessing different SIRETs:

Number of distinct known identifiers	1	2	3+	Total
Known SIRETs	21,791	432	156	22,548
	96.64%	1.92%	0.69%	100.00%
Known SIRENs	22,159	375	14	22,548
	98.27%	1.66%	0.06%	100.00%

Table 14. Distribution of the number of known identifiers by cluster, in terms of SIRETs and SIRENs.

SIRET	Name	Address	City
30059912300019	ASS NATIONALE POUR LA FORMATION PROFESSIONNELLE DES ADULTES	13 PL DU GENERAL DE GAULLE	MONTREUIL
30059912308228	ASS NATIONALE POUR LA FORMATION PROFESSIONNELLE DES ADULTES	13 PL DU GENERAL DE GAULLE	MONTREUIL

This, in turn, causes Dedupe to create clusters with homogeneous SIRENS but heterogeneous SIRETs.

Concluding Remarks These results show the performance of our clustering step in terms of clusters. But ultimately, we are interested in the quality of the agent SIRETs. From the perspective of false positives, we can distinguish three different outcomes. They depend on the SIRET assigned to each agent based on its cluster, according to the post-processing described in Section 5.3:

- *Full SIRET*: the SIRET is correct. This can match two situations: either the method puts the agent in a singleton cluster, or it puts it in a multi-SIRET cluster in which its SIRET is majority.
- *Partial SIRET*: the SIRET is incorrect, but the SIREN is correct. This happens when the method puts the agent in a multi-SIRET cluster in which the majority SIRET is different from the agent's, but with a common SIREN.
- *Incorrect*: neither the SIRET or the SIREN are correct. This situation corresponds to the case where the method puts the agent in a multi-SIRET cluster whose majority SIRET has nothing in common with the agent's.

Table 15 summarizes this aspect of the performance. A large portion of the clusters contain only a single known identifier and therefore, most agents are associated with the correct SIRET. Only 1.31% of the considered unique agents end up with a completely incorrect identifier, or 2.01% if we also include partially incorrect identifiers.

Number of distinct known identifiers	Full	Partial	Incorrect	Total
Count	23,762	169	393	24,324
Proportion	97.99%	0.70%	1.31%	100.00%

Table 15. Distribution of the three possible outcomes of the clustering process for the known SIRETs and SIRENs.

5.4.3 False Negatives

False negatives correspond to agents incorrectly placed in different clusters by Dedupe. In order to assess this type of error, we cannot use the same data as when studying the false positives in Section 5.4.2, because the forms taken by these agents do not exhibit enough diversity. Instead, we adopt a specific procedure to constitute a more appropriate ground truth.

First, we randomly sample the agent occurrences of the original TED dataset, under the following constraints. Each such occurrence must correspond to an agent appearing *several times* in the original TED data, and under *different forms* in this sample. Moreover, each sampled agent must have a SIRET[27]. Second, we ignore these agent occurrences during the identification step, which is only applied to the rest of the data. Third, we conduct the Dedupe-based clustering phase on the whole dataset, including the sample. Finally, we assess the false negatives produced during this last step by studying how the sampled agent occurrences that represent the same unique agent are distributed over the clusters identified by Dedupe.

Our sample contains $5{,}020$ occurrences, that correspond to 377 unique agents (i.e. there are 377 different SIRETs in the sample), for an average of 13.31 occurrences by agent. Each SIRET appears between 1 and 538 times in the sample. To assess how these occurrences are distributed over the Dedupe clusters, we compute two measures: the Concentration vs. Singleton Ratios.

Concentration Ratio The *Concentration Ratio* $CR(a)$ is defined for an agent a of interest. It is the maximal proportion of occurrences of this agent in a single cluster, relative to the total number of occurrences of this same agent in the sample:

$$CR(a) = \max_{C \in \mathcal{C}} \frac{|C \cap A|}{|A|}, \tag{2}$$

where $\mathcal{C} = \{C_1,...,C_k\}$ is the partition constituted of k clusters C_i detected by Dedupe, A is the set of all occurrences of a in the sample, and $|...|$ denotes the cardinality of a set. A concentration ratio close to one indicates that the occurrences of the same agent are located in a single cluster, and thus that the agent was well clustered by Dedupe. On the contrary, a low ratio shows that these occurrences are scattered over a number of clusters.

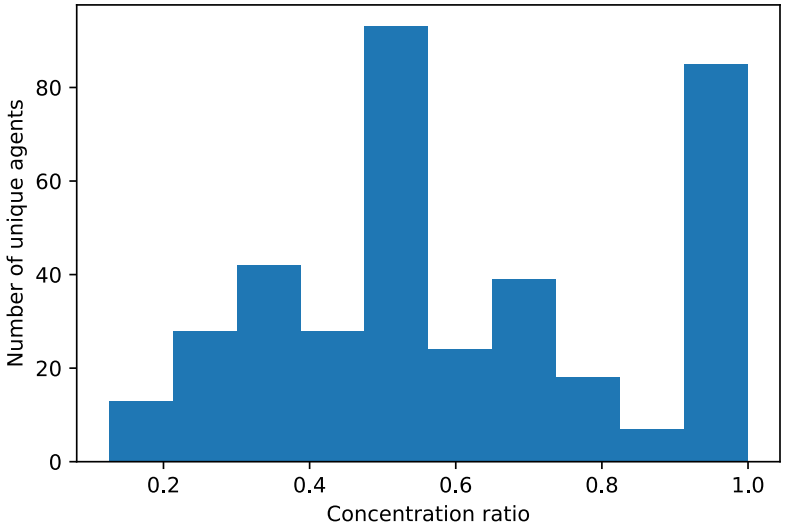


Figure 11. Distribution of the concentration ratio over unique agents.

Figure 11 shows the distribution of the concentration ratio over unique agents. The x-axis represents the concentration ratio, whereas the y-axis is the number of agents. The mean concentration ratio is 0.6, which means that Dedupe puts more than half of the occurrences of an agent in the same cluster, in average. Dedupe perfectly clusters 243 agent occurrences

[27] It can be a pre-existing SIRET, or a SIRET estimated at the identification step.

(5%), representing 83 unique agents (22%). These are clusters with 2 or 3 occurrences: it is apparently hard to gather many occurrences of the same agent in a single cluster. The others agents are less concentrated, with a majority of clusters containing half of their occurrences.

In order to define an overall performance measure, we sum the concentration ratio over all known SIRETs, using their frequencies as weights. We get a value of 0.43, which is consistent with our previous observations.

Singleton Ratio Among the occurrences that are not gathered in the same cluster, for a given agent, we consider differently those each constituting a singleton cluster, vs. those forming a cluster with some occurrences of other agents. Indeed, the former correspond to false negatives, whereas the latter are false positives. Since we focus on the former in this section, we propose the *Singleton Ratio* $SR(a)$ to characterize them. Like the *Concentration Ratio*, it is computed for an agent of interest a. It corresponds to the ratio of its number of occurrences forming singleton clusters, to its total number of occurrences in the sample:

$$SR(a) = \frac{|\{C \subset \mathcal{C} : C \subset A \wedge |C| = 1\}|}{|A|}. \tag{3}$$

A singleton ratio close to 1 indicates that all the agent occurrences are distributed in their own cluster. On the contrary, a ratio close to zero shows that the occurrences belong to larger clusters (possibly with other occurrences of the same agent, or not).

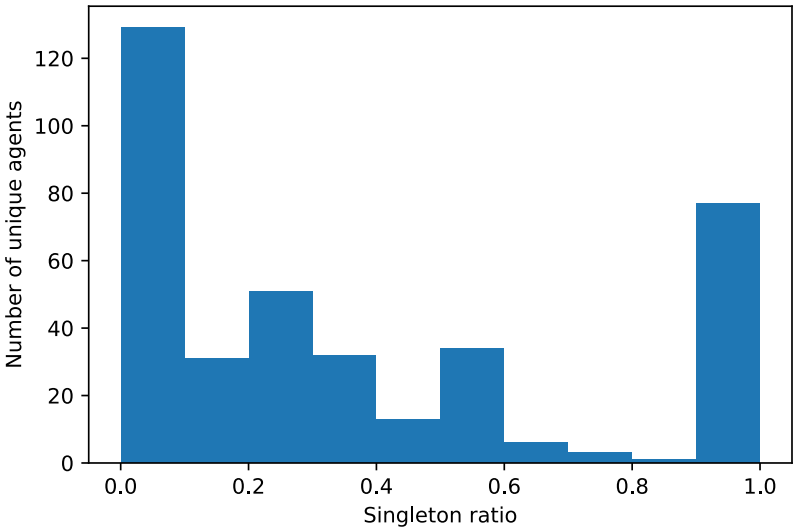


Figure 12. Distribution of the singleton ratio over unique agents.

Figure 12 shows the distribution of the singleton ratio over unique agents. The x-axis represents the singleton ratio and the y-axis the number of unique agents. On the one hand, there are 77 unique agents (20%) with a ratio of 1. This means that all related occurrences are located in separate clusters. This is the case for occurrences that are very different, for example occurrences not sharing the same name. This result confirms the relevance of our identification process: some occurrences can only be gathered by finding the correct SIRET number. On the other hand, the ratio of the rest of the agents is much lower. The mean singleton ratio for these other agents (i.e. without considering ratios equal to 1) is 0.18. This means that Dedupe splits the set of occurrences linked to a SIRET into a limited number of subgroups.

As with the CR before, we compute an overall measure by summing the singleton ratio over all agents, using their frequencies as weights. We get a value of 0.52 indicating a relatively high level of dispersion of the agent occurrences over the clusters.

Concluding Remarks As mentioned in Section 5.3, our post-processing includes the merging of agent occurrences located in the same cluster, even if they do not have the same SIRET. For this purpose, we keep the majority SIRET, i.e. the one associated to the largest number of occurrences in the cluster.

Considering a given unique agent at the end of the clustering step, some of its occurrences can be assigned the correct SIRET, but others could receive only a partially correct identifier (same SIREN), or a completely incorrect identifier, or even no identifier at all. Taking into account all possible outcomes makes it difficult to assess the performance of this step in a meaningful way. For this reason, we adopt a simplified view by focusing only on how the *absolute majority* of the agent's occurrences are treated. We distinguish the following five possible situations:

- *Full SIRET*: most of the agent's occurrences belong to the same cluster, whose majority SIRET matches the agent's. Most occurrences of this agent consequently get their correct SIRET.
- *Partial SIRET*: most of the agent's occurrences get a SIRET compatible with the agent's SIREN. This happens either when the occurrences are concentrated in the same cluster, but are a minority, or when the occurrences are scattered over several clusters.
- *Incorrect SIRET*: most of the agent's occurrences get a SIRET incompatible with the agent's SIREN. The situations leading to this case are similar to the previous one, except with completely different SIRETs.
- *No SIRET*: most of the agent's occurrences do not receive any SIRET at all.
- *No Decision*: there is no absolute majority for any of the four above situations. All of them may occur for the considered agent, but none dominates.

Table 16 summarizes this aspect of the performance. More than half of the unique agents (54%) see most of their occurrences left without any SIRET at all. This is because these occurrences are mostly located in singleton or unidentified small clusters. For 37% of the unique agents, most of their occurrences receive a SIRET. Focusing only on these cases, this SIRET is correct for 75% of the unique agents.

Unique agents	Full	Partial	Incorrect	None	No Decision	Total
Count	105	9	28	204	31	377
Proportion	27.85%	2.39%	7.43%	54.11%	8.22%	100.00%

Table 16. Distribution of the five possible outcomes of the clustering process.

To summarize these results: when our process is able to assign a SIRET to a unique agent, it is frequently correct; however our process is not able to identify a SIRET for most agents. This last comment should be modulated though, due to the data we used to perform this assessment. By construction, these data contain many occurrences of the same agent; however, in practice, this is unlikely, as it would require the agent occurrences to take a number of very different forms. The role of the identification step is to merge the distinct forms of the same unique agent that are similar enough. After this step, one agent should have either only one form, or a few very different ones.

5.4.4 Extrapolation

Assuming that the performance scores obtained at Sections 5.4.2 and 5.4.3 apply to the whole dataset, we can extrapolate to estimate the overall performance of our clustering step, in terms of both unique agents and agent occurrences.

Unique Agents Before the clustering step, we have 306,984 unique agents. Among them, 230,990 have a SIRET before this step, whereas 75,994 do not. The clustering process manages to find a SIRET for 2,243 of the latter, which means we are left with 73,751 unidentified unique

agents in the end. Table 17 summarize the situation.

Agent Type	Identified Before	Siretized Now	Not Siretized	Total
Unique agents	230,990	2,243	73,751	306,984
	75.24%	0.73%	24.02%	100.00%

Table 17. Status of the SIRET of *unique agents* after the clustering: SIRET already known before this step (first column), SIRETs estimated during this step (second column), and SIRETs that could not be estimated during this step (third column).

It is very difficult to assess the exact effect of the clustering process, because it involves not only the siretization of previously unidentified agents, as with the identification step, but also some merges likely to change the SIRET of certain already identified agents. On top of that, unidentified agents can also be merged together. Consequently, we must rely on some strong assumptions in order to extrapolate.

Agent Type	Correct	Incorrect	Total
Unique agents with a SIRET before the clustering step	189,957	41,033	230,990
	82.24%	17.76%	100.00%
Unique agents with a SIRET estimated through clustering	2,207	36	2,243
	98.38%	1.62%	100.00%

Table 18. Estimation of the numbers of unique agents with a correct vs. incorrect identifier. The top part concerns agents already possessing a SIRET before the clustering step. The bottom part focuses on those that acquire a SIRET through this step. Agents without any identifier at this stage are ignored in this table.

First, we assume that the siretization performance scores estimated for the set of unique agents *before* the identification step (presented in Table 11) also apply to the set of unique agents *after* this step. This allows us to estimate, among the 230,990 unique agents possessing a SIRET at this stage, how many have a correct vs. incorrect SIRET, as shown in the top part of Table 18. Both full and partial SIRETs are considered as correct. Second, we assume that the performance scores computed in Section 5.4.2 apply to the 2,243 agents that get a SIRET through the clustering process. This gives us the bottom part of Table 18.

Agent Type	Correct	Incorrect	Missing	Total
Unique agents	192,164	41,069	73,751	306,984
	62.60%	13.38%	24.02%	100.00%

Table 19. Extrapolation of the performance of the clustering step, in combination with the other steps of the proposed process, for *unique agents*.

We sum both parts of Table 18 and insert the agents without any SIRET to get the final values displayed in Table 19. It is worth stressing that the decrease in performance observed when comparing this table to Table 11 is only apparent. It is actually due to the strong decrease in number of unique agents between the identification and clustering steps. This, in turn, is caused by the many agents that can be merged thanks to the high performance of the identification step.

Agent Occurrences We proceed similarly to assess the overall performance in terms of agent occurrences instead of unique agents. Table 20 shows the distribution of occurrences depending on their status at the end of the identification step: correct SIRET, incorrect SIRET, or no SIRET at all.

We apply the same method as before to estimate the numbers of agent occurrences that are correctly/incorrectly identified during the clustering step. We get 2,204 correct

Agent Type	Identified Before	Siretized Now	Not Siretized	Total
Agent occurrences	2,674,915	2,240	201,528	2,878,683
	92.92%	0.08%	7.00%	100.00%

Table 20. Status of the SIRET of *agent occurrences* after the clustering: SIRET already known before this step (first column), SIRETs estimated during this step (second column), and SIRETs that could not be estimated during this step (third column).

SIRETs, whereas 36 occurrences get an incorrect SIRET. Summing with Table 20 yields the final results shown in Table 21. Unlike unique agents, the number of occurrences is constant throughout our process. Therefore, we do not have the same issue regarding the comparison of the proportions/numbers of correct/incorrect/missing SIRET between the different steps.

Agent Type	Correct	Incorrect	Missing	Total
Agent occurrences	2,313,622	363,533	201,528	2,878,683
	80.37%	12.63%	7.00%	100.00%

Table 21. Extrapolation of the performance of the clustering step, in combination with the other steps of the proposed process, for *agent occurrences*.

6 Conclusion and Perspectives

In this section, we first summarize the outcome of the process described in Sections 3–5, and discuss the main statistics of the resulting database (Section 6.1). We then briefly discuss *Opentender*, an alternative public procurement database, and compare it to our own (Section 6.2). Finally, we turn to the limitations still remaining in the current database, and propose some ways to solve them, essentially by leveraging additional secondary data sources (Section 6.3).

6.1 Process Outcome

In this section, we summarize the cost of the processing applied to improve the TED data and constitute the FOPPA database (Section 6.1.1). We also show and discuss how agent- and field completeness-related statistics describing the data evolve at each of the processing steps (Sections 6.1.2 & 6.1.3). Finally, we discuss statistics related to criteria (Section 6.1.4), and more generally to the whole database (Section 6.1.5).

6.1.1 Computational Cost

Table 22 shows the time required to perform each main step of our proposed process on an NVIDIA GeForce RTX 2080 Ti. The whole process is long mainly because of the identification step. However, in practice, we run this step in parallel on 10 such GPUs, therefore, it only takes 6 days, effectively.

Database Initialization Step	Identification Step	Clustering Step
1 hour	60 days	4 hours

Table 22. Time required for the main steps of the process.

The identification step is long because two costly operations occur at this step: on the one hand, the search among the large SIRENE database, and on the other hand, the matching of names and addresses on many possible candidates. In order to alleviate this, we use the department as a first filter. However, some places like Paris concentrate a very large number of agents, and thus force the script to make a lot of comparisons.

This process is designed to be performed only once, when the database is created. For this reason, we did not try to optimize our source code. Moreover, the whole database will be published online, and publicly available, so no one else will have to perform it again.

6.1.2 Number of Agents

Table 23 represents the number of unique agents at each step of the process. When an agent appears both as a buyer and a winner, we use the majority rule to put it in a single category.

The first row (*Raw Data*) shows the number of unique agents at the beginning, i.e. before any processing. We just compare exactly the strings present in each field associated with an agent, i.e. the name, SIRET, address, city and zipcode. The number of unique buyers is much lower than the number of unique winners. This is because buyers fill more carefully their own information than they do with the winners'. Therefore, more of the buyer occurrences exactly match, and can be associated to the same unique agent.

The second row (*After Separation*) shows the situation after having separated multiple agents listed in the same lot. Again, we directly compare the description fields to identify unique agents. This separation increase the number of buyers by 13%, with $18{,}142$ new entries

Step	Buyers	Winners	Total
Raw Data	122,577	679,883	802,460
After Separation	140,719	794,015	934,634
After Normalization	98,574	633,790	732,364
After Identification	28,032	278,952	306,984
After Clustering	26,618	274,478	301,096

Table 23. Number of unique agents counted originally, and after each step of the proposed process.

and 14% for winners, with 114,032 new entries. It is more common to have several winners for a single lot than several buyers, which explains the superior increase on the winner side.

The third row (*After Normalization*) shows the number of unique agents after the normalization of their names. During this step, we gather similar occurrences of agents with a few differences in punctuation and typography. This is an important step, since it reduces the number of buyers by 30% and the number of winners by 20%.

The fourth row (*After Identification*) focuses on the situation after the identification step. During this step, we gather occurrences representing the same agent, but filled with dissimilar information. It is the main step of our processing, as it allows gathering agent occurrences that take relatively different forms, unlike the previous steps.

The fifth row (*After Clustering*) describes the agents after the clustering step, i.e. at the end of the process. This step is more conservative than the previous one, and only merges occurrences with very similar forms. It aims at detecting occurrences that are too poorly filled in order to be siretized before, but close enough to know that they represent the same agent.

6.1.3 Field Completeness

Table 24 shows the level of completion of the agent fields in the original data (*Before* column) and after our processing (*After* column). For each field, the table exhibits the proportion of lots in which it is filled (whatever the content), separately for the buyers and winners.

Field	Buyers		Winners	
	Before	After	Before	After
Name	100.0%	100.0%	91.7%	100.0%
SIRET	16.4%	97.5%	2.9%	86.2%
Address	98.8%	99.3%	72.5%	89.1%
City	100.0%	100.0%	82.6%	90.0%
Zipcode	98.7%	99.3%	79.0%	88.7%

Table 24. Completion of the main fields that describe agents.

The increase is important for all fields, but more particularly for the SIRET, due to our efforts during the identification step. In addition to the SIRET itself, this step also allows retrieving some information that covers the other fields: name, address, city, zipcode. This, in turn, allows completing the missing values, thus increasing the completion rate. In addition, this also allows correcting or unifying the values already present in the database. The clustering step also has an effect, albeit weaker. Indeed, some of the agents that it gathers are poorly filled: the resulting merged agents generally still have missing fields.

Let us focus on the SIRET field, which is the most important. In terms of agent occurrences, after the separation step, and discarding unsuccessful tenders (because they have no winner), we have 1,497,414 buyers and 1,381,269 winners, for a total 2,878,683 occurrences. These

values are constant for the rest of the process: only the ratio of identified to unidentified agents evolves at each step of our pipeline. Table 25 shows this evolution for each remaining step. Both *Present* columns account for agent occurrences with a SIRET, whereas both *Absent* columns deal with the agent occurrences with a structurally invalid SIRET or no SIRET at all.

Step	Buyers		Winners	
	Absent	Present	Absent	Present
After Separation	1,290,098	207,316	1,341,235	40,034
After Normalization	1,019,993	477,421	1,209,520	171,749
After Identification	24,248	1,473,166	179,520	1,201,749
After Clustering	23,226	1,474,188	178,302	1,202,967

Table 25. Number of agent *occurrences* with a SIRET (*Present* columns), or with a structurally invalid SIRET or no SIRET at all (*Absent* columns), after each step of the proposed process. The total is constant, at 2,878,683.

As expected, the strongest effect is clearly due to the identification step, which allows identifying the SIRET of 2,025,745 agent occurrences, representing 90.87% of the SIRETs missing before this step. By comparison, the clustering step allows retrieving the SIRET of only 2,240 agent occurrences, i.e. 1.10% of the unidentified occurrences remaining after the previous step. The same observation holds when considering unique agents, see Table 26.

Step	Buyers		Winners		Agents		
	Absent	Present	Absent	Present	Absent	Present	Total
After Separation	124,143	16,576	678,894	28,375	803,037	44,951	847,988
After Normalization	83,982	14,592	606,447	27,343	690,429	41,935	732,364
After Identification	2,278	25,754	73,716	205,236	75,994	230,990	306,984
After Clustering	1,877	24,741	71,874	202,604	73,751	227,345	301,096

Table 26. Number of *unique* agents with a SIRET (*Present* columns), or with a structurally invalid SIRET or no SIRET at all (*Absent* columns), after each step of the proposed process. Unlike with the agent *occurrences*, the total evolves at each step.

A manual analysis of the agents that our process cannot identify, or identifies incorrectly, reveals that they correspond to situations where agent information is so incomplete and/or incorrect that even handling them manually proves to be a very difficult, or even impossible, task. For instance, 52,598 (71%) of the unique 73,751 agents remaining unidentified at the end of the process have no location information whatsoever.

6.1.4 Award Criteria

Of the 1,380,965 lots, 1,041,242 (78%) contain some information regarding the award criteria. Thanks to our criteria processing, the 1,041,242 raw strings originally present in the data to describe these criteria are separated into 2,910,408 criteria, and each one is associated to a coarser class (*Price*, *Technique*, *Delay*, *Social*, *Environmental*, *Other*).

Price	Technique	Delay	Social	Environmental	Other	Nothing
989,403	888,295	156,669	26,326	137,875	227,745	298,670
71.65%	64.32%	11.34%	1.91%	9.98%	16.49%	21.63%

Table 27. Distribution of criterion classes over lots. Several criteria can be used in one lot.

Table 27 shows the distribution of these criterion classes over the concerned lots. Column *Nothing* corresponds to the lots without any specified criterion. Note that a lot can rely on

several criteria at once. Most of the lots use criteria related to price or technique.

6.1.5 General Statistics

In the end, the FOPPA database contains 1,380,965 lots, which are described by 410,283 CANs. Among them, 286,160 are linked to a CN, and are therefore complemented with some additional information, such as the publicity duration. They involve 301,096 unique agents. Table 28 indicates the number of entries in each data table in the FOPPA database.

Database Table	Number of entries
`Lots`	1,380,965
`Criteria`	2,910,408
`Agents`	301,096
`Names`	506,061
`LotBuyers`	1,497,632
`LotSuppliers`	1,371,535

Table 28. Size of each table in the FOPPA database, at the end of the processing.

6.2 Comparison with Opentender

The Opentender [10] database[28] is very similar to ours, so it is worth comparing both of them in order to better highlight their similarities and differences. It was created in the framework of the H2020 Digiwhist project[29] (2015–2018), which aimed at identifying public sector corruption. Opentender is maintained by a consortium constituted of academic institutions, companies, and non-for-profit organizations[30].

In this section, we first describe the dataset and summarize how we match it to the FOPPA (Section 6.2.1). We then discuss the differences between both datasets (Section 6.2.2).

6.2.1 Description

Opentender provides access to the public procurement data of 35 European countries, as well as to some data describing the involved economic agents. We first present briefly the characteristics of this dataset, before describing the method we use to match its objects to FOPPA objects.

Sources According to the Digiwhist technical report dedicated to raw data [12], Opentender is based on three categories of data sources. The first are *public procurement databases*, including the TED and 20 national databases. The Digiwhist consortium mainly scrapped the Web versions of these databases, with a few CSV dumps in addition. For France, the national source is the BOAMP[13, p.27], which we introduce in Section 1.1.

The second category concerns *company data*, including registry information (name, identifier, date of creation, address, etc.), financial information (annual turnover, profit rate, etc.), ownership and manager information. The Digiwhist consortium purchased these data from Bureau van Dijk, a private data provider.

The third category is about *public sector data*, which include contracting authorities, public officials, and budgets. The consortium leveraged a heterogeneous collection of public documents for this purpose: websites, databases, PDF documents, and others. Due to this heterogeneity, and to the fact that not all documents are available for all countries, they could not cover certain countries, including France.

[28] https://opentender.eu/
[29] https://digiwhist.eu/
[30] https://opentender.eu/fr/about/about-opentender

Organization The data themselves are available under two forms. First, CSV and JSON dumps, that list contract and award notices between 2009 and 2022. Second, an online version, which includes the information associated with a notice/buyer/winner, as well as a dashboard providing different indicators and redflags.

In the CSV File, each line represents an award between a buyer and a winner. The complete list of variable is available online[31]. Therefore, a lot can be represented over several lines, provided more than two agents are involved (several buyers or several winners). This is a major modeling difference with FOPPA: here, each row is a contractual relationship between a buyer and a winner, whereas in FOPPA, each row represents a lot, which can involve several buyers or winners.

Another important difference is the representation of tenders. In Opentender, they are identified by a unique identifier, which gathers all notices associated to the tender, i.e. possibly several award notices, and they can come from both the TED and a national database. However, the Opentender data model only stores the URL of the very last source document included (e.g. BOAMP or TED notice). Consequently, it is not possible to retrieve all the source documents related to a given tender: this complicates matching Opentender and FOPPA objects, as the latter consistently uses TED identifiers.

Matching Method Because Opentender and FOPPA are not compatible in terms of how they model lots and award notices, it is not possible to directly compare them: we need a specific process to match them.

On the one hand, we parse the Opentender tenders and select those associated to a TED URL (by opposition to a BOAMP URL)[32]. We keep only the entries published during the 2010–2020 period, and extract their TED identifiers from the URLs. On the other hand, we retrieve the notices associated to these identifiers in the FOPPA.

We are able to retrieve $254{,}075$ award notice identifiers available on both sides. This represents 62% of all the award notices available in FOPPA. They describe $924{,}895$ lots involving $1{,}011{,}391$ buyers and $923{,}176$ winners in FOPPA. By comparison, in Opentender, they correspond to $843{,}522$ contractual relations between a buyer and a winner. This discrepancy is explained later in Section 6.2.2.

6.2.2 Comparison

Opentender is clearly more ambitious than FOPPA, in terms of geographic coverage (35 states vs. 1) and attribute span (public procurement, individual and budgetary data vs. only procurement data). However, as stressed by the consortium itself in its technical report focusing on data validation [19], elaborating such a database is a real challenge:

> "[...] it takes years of work to actually fine-tune data extraction even from a single procurement source. [...] Based on this experience, Digiwhist's goals are ambitious, even if we only aim to achieve moderate quality data within the project. Note that since the Digiwhist team aims to further improve the quality of the data during the sustainability period, the data quality results are to be taken rather as state-of-the-art."

Indeed, when comparing the part of Opentender that can be matched to the FOPPA, we detect several issues. Most of them are shared with the FOPPA, and are already discussed before in this report. In the rest of this section, we focus on the issues present in Opentender but not in the FOPPA.

Missing Lots The first issue concerns the absence of certain lots in Opentender, compared to the TED and FOPPA. As an example, we consider award notice number 643262-2020, which

[31] https://drive.google.com/file/d/1fb7CgXJ2dqpbYujZRF8RU0gEBHwOojGM/view

[32] We leveraged field `tender_publications_lastContractAwardUrl` of Opentender.

can be found in both Opentender[33] and the TED[34] (and consequently FOPPA). Figure 13 provides screen captures of both Web pages. Opentender associates this notice to a single lot, when there are 66 of them according to the TED (as indicated in yellow in the figure). The CSV version of Opentender also lists a single lot for this notice. A search for the notice title with the Opentender search engine does not return any additional match.

Figure 13. Screen captures for the example of missing lots between FOPPA (left) and Opentender (right). The notice identifier and the numbers of lots are highlighted in green and yellow, respectively.

When considering the whole set of $254{,}075$ matching notices identified in Section 6.2.1, we detect the same problem for $34{,}944$ of them (13.8%). Note that there are $410{,}283$ notices in the FOPPA in total, so if we extrapolate to the rest of the database, this number becomes $56{,}619$. In average, each notice of the FOPPA is associated to 3.37 lots. Therefore, we estimate the number of missing lots to be approximately $190{,}806$ in total (14%).

Agent Mismatches Opentender does not store the national identifiers of the agents (the SIRET, in our case). Instead, the database relies on an internal identifier assigned after a deduplication process that relies on the same principle as ours [13, p.50]. The process starts with the standardization of agent names and addresses. A manual matching step allows human operators to indicate that certain agents are occurrences of the same entity. An exact matching step focuses on the name, address and all available identifiers to gather identical occurrences of the same agent. It is complemented by an approximate matching step, that leverages the name, address, zipcode and NUTS code (Nomenclature of Territorial Units for Statistics[35]) to gather occurrences not yet grouped at the previous steps. In order to reduce the computational load, the process groups occurrences into small groups [13, p.52] and makes internal comparisons. The problems encountered during these parts are similar to those that we describe in Section 2.3.7.

Database	Unique buyers	Unique winners	Total
Opentender	$48{,}654$	$413{,}819$	$462{,}473$
FOPPA	$23{,}268$	$204{,}792$	$228{,}060$

Table 29. Number of unique agents in Opentender and the FOPPA for the $254{,}075$ matching notices.

[33] https://opentender.eu/fr/tender/553de9cb-1acd-478d-a4be-c3ef3d3c2911
[34] https://ted.europa.eu/udl?uri=TED:NOTICE:643262-2020:TEXT:EN:HTML
[35] https://ec.europa.eu/eurostat/web/nuts/background

In order to compare Opentender and the FOPPA, we look at their numbers of unique agents for the subset of $254{,}075$ notices described in Section 6.2.2, as shown in Table 29. There are roughly twice as many agents in Opentender as in the FOPPA, which is a significant difference.

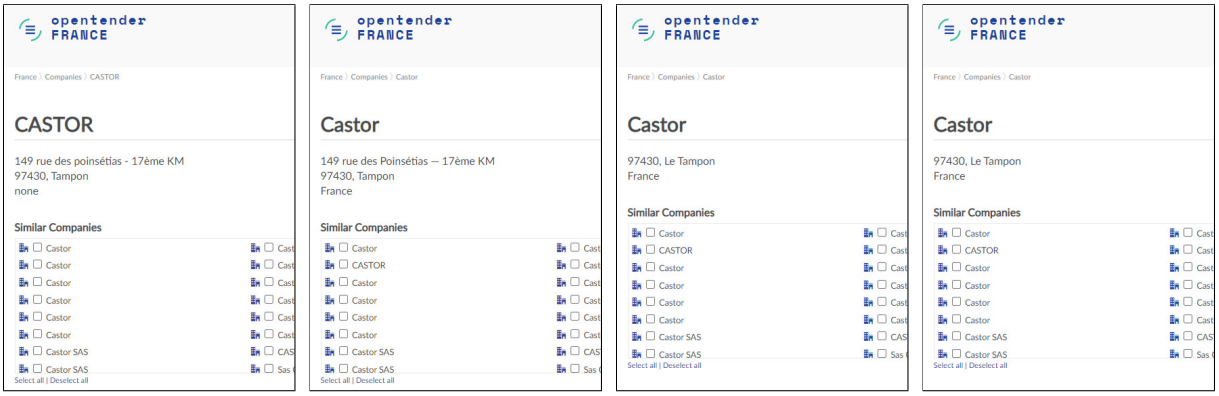


Figure 14. Four distinct Opentender agents that actually correspond to the same entity. The Opentender page of the first one lists 20 other occurrences that seem equally similar.

When browsing the page of an agent on the Opentender Website, the GUI displays a number of similar (but distinct) agents. This quickly reveals a number of very similar agent occurrences, that should visibly be merged, but are not. This indicates that the larger number of unique agents in Opentender can be explained by missed matches. Figure 14 shows an example of four such agents[36] that differ only on minor details, and are modeled as a unique agent in the FOPPA.

Spurious Winners As explained in Section 2.3.1, when a tender ends up unsuccessful, this situation must be stated explicitly through a contract award notice. Obviously, there is no winner in this case, and this is indicated by filling the winner's name field with some text describing the situation, typically `infructueux` (French for unsuccessful), `sans suite` (dismissed), or `aucune offre retenue` (no tender accepted). However, in Opentender, these indications are considered as proper economic agents, resulting in spurious entries such as `Sans suite` [37], `Infructueux` [38], and `néant` [39]. Moreover, due to the agent mismatch issue discussed in the previous paragraph, the same string can be associated to several, and sometimes many, distinct unique winners.

We identified a set of strings likely to describe an unsuccessful procedure, and counted the number of concerned winners in Opentender. As before, to be consistent with the previous results, we focus only on the $254{,}075$ matching notices identified in Section 6.2.1. Table 30 shows the results of this verification. A total of $18{,}430$ unique winners exhibit this issue (4.45%). By comparison, the FOPPA does not contain such spurious entries.

Missing Criteria Like the TED and FOPPA, Opentender provides the criteria selected to award the lots. However, this information is not always present, even when it is in the TED.

[36] https://opentender.eu/fr/company/FR_body_5b2e901209c37ca2fd8232e443b36d8a4cc24ec8af975247eb1d770c072ad0f9, https://opentender.eu/fr/company/EU_body_92a09aaf5281975b5b5ee179325bd94510dbef51add018cb78a20ebd28b650cb, https://opentender.eu/fr/company/EU_body_0f3327902956e78c0c55e6e6ae9854d40d9a5ef3ca692074208c218a2adb70b0, https://opentender.eu/fr/company/EU_body_9f4fdbb1f3ec7a68b27771bc8b928d0c604a1263bdf6cb77de7cdb6bf3fddb51

[37] https://opentender.eu/fr/company/EU_body_cf5995de70c28f04b99d5a7d3079afa35010d47734c89d31fa09e60c5ca8d9b6

[38] https://opentender.eu/fr/company/EU_body_010e9b539df8963062c6fc99b0c55fcf30b3137f956b55bab9fb8453f2d75145

[39] https://opentender.eu/fr/company/EU_body_c8b1b9d36df6b1d9301d8ec048d77e995abdedda8a9e6918971f167c6ce49620

Name	Meaning	Count	Proportion
`Infructueux`	Unsuccessful	10,196	2.46%
`Sans suite`	Dismissed	7,197	1.74%
`Non attribué`	Not attributed	262	< 0.01%
`Information non connue`	Unknown information	258	< 0.01%
`Aucune offre`	No offer	95	< 0.01%
	Empty or blank string	171	< 0.01%
`Abandon`	Withdrawal	130	< 0.01%
`Néant`	Void	121	< 0.01%
Total		18,430	4.45%

Table 30. Number of spurious unique winners corresponding to unsuccessful procurement procedures in Opentender. The proportions are relative to the number of unique winners in the selected notices, i.e. 413,819.

This is for example the case with award notice number 124665-2017, present in the TED[40] and Opentender[41]. As shown in Figure 15, the former lists two criteria (price and technical value), whereas the place that should indicate these criteria is empty in the Opentender page. In the CSV version, there is a field that counts the number of criteria for a notice: in this case, it contains the value `NULL` value, which is consistent with the web page, but not with the TED data. On the contrary, these criteria are available in the FOPPA.

Figure 15. Screen captures for the example of missing criteria between the TED (left) and Opentender (right). The notice identifier is highlighted in green, as before. In the TED page, the orange color highlights the specified criteria, whereas in the Opentender page, it indicates the empty space that they should fill would they be present.

When considering the whole set of 254,075 matching notices identified in Section 6.2.1, we identify the same problem (lots in the TED data that are missing on Opentender) in 91,566 notices (36.0%). By extrapolating to the 410,283 notices of the FOPPA in total, this issue concerns 147,862 notices associated to an estimate of 498,294 lots (36.1%).

6.3 Possible Improvements

The current FOPPA database can be improved in two ways. First, some of the problems identified in Section 2.3 are still partially present, and require some additional processing to be fixed (Section 6.3.1). Second, the informative value of the FOPPA could be increased by integrating some additional information coming from certain secondary sources of interest (Section 6.3.2).

[40] https://ted.europa.eu/udl?uri=TED:NOTICE:124665-2017:TEXT:FR:HTML
[41] https://opentender.eu/fr/tender/5d66e3bb-ad3e-4d0b-9129-d352a13bcd29

6.3.1 Solving the Remaining Problems

In this section, we review the open problems remaining in the FOPPA dataset, and the potential ways to solve them.

One potential solution to all open problems is to leverage alternative types of access to the TED data. In this work, we have only used the CSV files so far, but as explained in Section 2.1.1, it is also possible to retrieve some information from the TED by two other means: first through an API; and second under the form of Webpages. It is not clear as of now if they allow accessing additional information compared to the CSV files, but if they do, this would allow us completing the FOPPA database.

Missing Notices This problem, identified in Section 2.3.1, is still open, as we did not try to tackle it when designing the process described in this report.

One possible solution would be to leverage the BOAMP[42] (*Bulletin Officiel des Annonces des Marchés Publics* - Official bulletin for public procurement notices), which is the national equivalent of the TED for France. Like for the TED, the BOAMP data are publicly available online. It covers a wider range of French notices, as it host not only contracts above the European threshold, but also lesser contracts that are below this threshold (but still above a –lower– national threshold).

Consequently, the BOAMP is supposed to subsume the TED, at least for French contracts. In practice, public procurement notices are first sent to the BOAMP, and then fetched to the TED if they are above the European threshold. Thus, one could assume that for a given such notice, the information available on the TED and the BOAMP are exactly the same. However, our examination of the available data revealed that this is not always the case. This means that the BOAMP data could be used to supplement the TED data, in particular regarding the missing CN and/or CAN.

Missing Agent Identifiers The absence of many agent identifiers is one of the most serious problem of the TED dataset, as explained in Section 2.3.7. Our process allows retrieving a lot of them, but does not completely solve this issue. We identify four potential solutions to find the remaining missing SIRETs and therefore complement the FOPPA database.

First, the most direct solution consists in studying the cases for which our method is not able to retrieve a SIRET at the identification step. This implies manually looking for the agent in the SIRENE database, so this task could be quite long. This analysis could help us propose new automatic methods covering cases that are ignored in the current version of our process.

Second, another straightforward solution consists in improving the merging of agents after our clustering step. We plan to define an additional step for this purpose, based on structural similarity in the buyer-winner graph. In this graph, vertices represent agents, and edges model contracts between them. Two vertices are structurally similar when they have the same (or almost the same) neighbors. In public procurement terms, this means that two structurally similar agents concluded contracts with the same other agents. Therefore, if one of them is unidentified, they are likely to be two occurrences of the same agent, and could thus be merged in our database.

The other solutions require leveraging additional data sources. The third potential solution makes use of the VIES (VAT Information Exchange System) to identify agents in the FOPPA database, instead of the national identifier that we currently use. The VIES is a code used at the European level to trace firms that have a commercial activity spanning several member states. However, this task seems difficult to implement due to the way the VIES database is managed. It is not a centralized European database: each member state is in charge of handling their own agents. In doing so, each state is likely to apply its own rules. In particular, agents that do not have any transborder commercial activity for some duration (which depend on the concerned member state) are automatically removed from

[42] https://www.boamp.fr/pages/entreprise-accueil/

the VIES database [7]. Therefore, it seems difficult to use this resource to retrieve any historical information, i.e. to process most of the contracts in the FOPPA. Finally, there is no direct access to the database content: it is only possibly to very the existence of a VAT number online[43].

Fourth, it is possible to leverage various commercial sources. The Website Societe.com[44] provides data regarding private companies, their director, interconnections, etc., for all companies listed in the RCS[45] (*Registre du Commerce et des Sociétés* – Register of Commerce and Companies). The Website Infogreffe.fr[46], which manages the RCS, provides legal information regarding the companies. This service is not fully commercial, as a part of their data are open. The Website Pappers.fr[47] seems to give access to an improved and revised version of the same data.

Missing Prices ...TODO...

...Other... ...TODO...

6.3.2 Extending the DB Perimeter

In this section, we discuss how the FOPPA can be extended in order to include more data, and in particular some additional types of entities.

DECP In France, each public authority must publish, for contracts over €25,000, certain data, known as DECP (*Données Essentielles de la Commande Publique* – Essential public procurement data). The DECP database[48] contains some information describing contract awards, more precisely:

- Contract identifier;
- SIRET of the buyer;
- SIRET of the winner;
- Description of the contract;
- Type of the contract (framework agreement, etc...);
- Publication date;
- Notification date;
- Price of the contract;
- CPV code;
- Duration of the contract;
- Localization of the contract;
- Procedure used (open procedure, etc.);
- Every further corrections.

Combining this database with FOPPA would then add some new contract award notices, and may help us to fill some missing data.

BRÉF The BRÉF database [14] (*Base de données Révisée des Élu·es de France* – Revised database of elected representatives of France) contains information about all persons holding an elected position in France during the fifth Republic, i.e. since 1958. It is not publicly available yet, but some DeCoMaP researchers participate in its constitution, so we could use these data if needed.

Each elected representative holds a sit at a public institution, which holds a SIRET. It is therefore possible to connect agents from the FOPPA (municipalities, departmental and

[43] https://ec.europa.eu/taxation_customs/vies/

[44] https://www.societe.com/

[45] https://www.economie.gouv.fr/entreprises/registre-commerce-societes-rcs

[46] https://www.infogreffe.fr/

[47] https://www.pappers.fr/

[48] https://www.data.gouv.fr/en/datasets/donnees-essentielles-de-la-commande-publique-fichiers-consolides/

regional councils, national assembly and senate) to some elected individuals from the BRÉF.

INPI The INPI database[49] (*Institut National de la Propriété Industrielle* – National institute of industrial property) is the French public service handling patents, brands, and industrial property. It is also in charge of providing the open data associated to this service, called the RNE[50] (Registre National des Entreprises – *National register of companies*). The INPI provides a dedicated online API[51] to access these data, which include the identity of the person representing the companies. This database could allow connecting the FOPPA with individuals involved in the private sector.

BODACC The BODACC database[52] (*Bulletin Officiel des Annonces Civiles et Commerciales* – Official Bulletin of Civil and Commercial Notices) includes all acts written in the RCS. In particular, it provides some information on:

- Sales and transfers of companies;
- Changes and deletions of natural or legal persons.

Combining this database with FOPPA would then make it possible to connect companies with individuals.

BANATIC The BANATIC database[53] (*Base nationale sur l'intercommunalité* – National database of intercommunal structures) is a database managed by the French ministry of Interior. It contains information regarding the nature and functioning of the French administrative structures that gather several municipalities, and consequently that handle a number of public procurement markets. These data could help better connect the FOPPA to the BRÉF.

Others The *Annuaire des Entreprises*[54] (Directory of Companies) is a public service that gathers information regarding public and private companies, by leveraging many public services: ministries, INPI, INSEE, VIES, and others.

The DINUM[55] (*Direction Interministérielle du Numérique* – Inter-ministerial directorate for digital services) also proposes an API[56] dedicated to the retrieval of data describing French companies.

The OFGL[57] (*Observatoire des Finances et de la Gestion publique Locales* – Observatory of local finance and public management) proposes an online service[58] to access the data it generates through its monitoring activity.

[49] https://data.inpi.fr/
[50] https://registre.entreprises.gouv.fr/
[51] https://data.inpi.fr/
[52] https://www.bodacc.fr/
[53] https://www.banatic.interieur.gouv.fr/
[54] https://annuaire-entreprises.data.gouv.fr/
[55] https://www.numerique.gouv.fr/dinum/
[56] https://api.gouv.fr/les-api/api-recherche-entreprises
[57] https://data.ofgl.fr/
[58] https://data.ofgl.fr/pages/accueil/

References

[1] R. Ackermann, M. Sanz, and A. Sanz. *Gaps and Errors in the TED database*. Technical Report PE 621.804. European Parliament's Committee on Budgetary Control, 2019. URL: https://www.europarl.europa.eu/RegData/etudes/IDAN/2019/621804/IPOL_IDA(2019)621804_EN.pdf (visited on 11/19/2022).

[2] BOAMP. *Avant de répondre à un marché public*. Bulletin Officiel des Marchés Publics. 2020. URL: https://www.boamp.fr/Espace-entreprises/Comment-repondre-a-un-marche-public/Questions-de-reglementation/Avant-de-repondre-a-un-marche-public/Sommaire (visited on 10/07/2021).

[3] C. Csáki. "Towards Open Data Quality Improvements Based on Root Cause Analysis of Quality Issues". In: *International Conference on Electronic Government*. Vol. 11020. Lecture Notes in Computer Science. Springer, 2018, pp. 208–220. DOI: 10.1007/978-3-319-98690-6_18.

[4] C. Csáki and E. Prier. "Quality Issues of Public Procurement Open Data". In: *International Conference on Electronic Government and the Information Systems Perspective*. Lecture Notes in Computer Science. Springer, 2018, pp. 177–191. DOI: 10.1007/978-3-319-98349-3_14.

[5] Directorate-General for Internal Market, Industry, Entrepreneurship and SMEs (GROW). *TED CSV Open Data – Advanced Notes on Methodology*. Tech. rep. Tenders Electronic Daily, 2020. URL: https://data.europa.eu/euodp/repository/ec/dg-grow/mapps/TED_advanced_notes.pdf.

[6] Directorate-General for Internal Market, Industry, Entrepreneurship and SMEs (GROW). *TED CSV Open Data – Notes & Codebook Version 3.4*. Tech. rep. Tenders Electronic Daily, 2021. URL: https://data.europa.eu/euodp/en/data/storage/f/2022-02-14T122429/TED(csv)_data_information_v3.4.pdf.

[7] European Union. *Check a VAT number (VIES)*. European Union. 2022. URL: https://europa.eu/youreurope/business/taxation/vat/check-vat-number-vies/index_en.htm (visited on 12/03/2022).

[8] European Union. *Statistical Report 2014, 2015 and 2016 on Public Procurement According to the Agreement on Government Procurement*. Communication. European Union, 2020. URL: https://members.wto.org/crnattachments/2020/GPA/EEC/20_1435_00_e.pdf.

[9] M. Fazekas and I. J. Tóth. "From corruption to state capture: A new analytical framework with empirical applications from Hungary". In: *Political Research Quarterly* 69.2 (2016), pp. 320–334. DOI: 10.1177/1065912916639137.

[10] Government Transparency Institute. *opentender.eu dataset*. 2018. URL: https://opentender.eu/all/download (visited on 03/18/2023).

[11] R. W. Hamming. "Error Detecting and Error Correcting Codes". In: *Bell System Technical Journal* 29.2 (1950), pp. 147–160. DOI: 10.1002/j.1538-7305.1950.tb00463.x.

[12] J. Hrubý, M. Fazekas, T. Pošepný, J. Krafka, M. Říha, M. Mikeš, and T. Mrázek. *D2.4 - Raw data*. Technical Report. Digiwhist, 2016. URL: http://digiwhist.eu/wp-content/uploads/2016/06/D2.4-final.pdf.

[13] J. Hrubý, T. Pošepný, J. Krafka, B. Toth, and J. Skuhrovec. *D2.8 - Methods Paper*. Technical Report. Digiwhist, 2018. URL: http://digiwhist.eu/wp-content/uploads/2018/03/D2.8-revised-version-FINAL.pdf.

[14] V. Labatut, N. Févrat, and G. Marrel. *BRÉF – Base de données Révisée des Élu·es de France*. Technical Report. Avignon Université, 2020. URL: https://hal.archives-ouvertes.fr/hal-02886580.

[15] V. I. Levenshtein. "Binary codes capable of correcting deletions, insertions, and reversals". In: *Soviet Physics Doklady* 10.8 (1966), pp. 707–710. URL: https://nymity.ch/sybilhunting/pdf/Levenshtein1966a.pdf.

[16] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008. DOI: 10.1017/cbo9780511809071.

[17] L. Potin, V. Labatut, R. Figueiredo, C. Largeron, and P.-H. Morand. *FOPPA: a database of French Open Public Procurement Award notices*. Tech. rep. Avignon Université, 2022. URL: https://hal.archives-ouvertes.fr/hal-03796734.

[18] L. Potin, V. Labatut, C. Largeron, and P. H. Morand. "FOPPA: an open database of French public procurement award notices from 2010–2020". In: *Scientific Data* 10 (2023), p. 303. DOI: 10.1038/s41597-023-02213-z.

[19] J. Skuhrovec, M. Říha, and M. Palanský. *D2.7 - Data validation results*. Technical Report. Digiwhist, 2018. URL: http://digiwhist.eu/wp-content/uploads/2018/03/D2.7-revised-version-FINAL.pdf.

[20] R. R. Sokal and J. F. Rohlh. "The comparison of dendrograms by objective methods". In: *Taxon* 11 (1962), pp. 33–40. DOI: 10.2307/1217208.

[21] M. Vingron and M. S. Waterman. "Sequence alignment and penalty choice". In: *Journal of Molecular Biology* 235.1 (1994), pp. 1–12. DOI: 10.1016/s0022-2836(05)80006-3.

A Database Changelog

Table 31 describes the different versions of the FOPPA database.

Version	Date	Description
`1.0.0`	2022/01/26	First version of the FOPPA database.
`1.0.1`	2022/03/28	Five changes: 1) Added Boolean fields to `Lot` table; 2) Fixed issues related to criterion weights; 3) Fixed normalized issues related to missing data; 4) Deleted agents related only to non-awarded lots; 5) Normalized dates.
`1.0.2`	2022/04/22	Added two new fields: `country` (country of the agents); `department` (additional information for overseas departments and Corsica).
`1.0.3`	2022/05/31	Correction of non-siretized agent names and geographical information.
`1.0.4`	2022/10/19	Added four new fields: `longitude` and `latitude` (position of the agents), `duration` (duration of the framework agreement) and `publicityDuration` (time allowed to make an offer).
`1.1.0`	2022/11/29	1) Added the 2020 data; 2) Fixed issues with certain fields names.
`1.1.1`	2022/12/15	1) Fixed incorrect types declaration in the SQL dump; 2) Fixed issues with certain field names.
`1.1.2`	2023/02/21	Change of terminology from *client* to *buyer*.
`1.1.3`	2024/03/07	1) Corrections of a few typos in the SQL script; 2) Modification of N/Y values to 0/1 in Boolean fields.

Table 31. History of the FOPPA database versions.

B Procedure-Related Information

As explained in Section 1, when a contract is above the European threshold of the concerned activity domain, it is necessary to follow a formalized procedure. Table 32 shows the evolution of the European thresholds over time.

Buyer	Sector	Type of contract	Threshold (€)
01/01/2004–31/12/2004			
Central public authority	All	Goods, Services	162,000
Local public authority	All	Goods, Services	249,000
Public entity	All	Goods, Services	499,000
Public authority/entity	All	Works, Concessions	6,242,000
01/01/2005–31/12/2005			
Central public authority	All	Goods, Services	154,000
Local public authority	All	Goods, Services	236,000
Public entity	All	Goods, Services	473,000
Public authority/entity	All	Works, Concessions	5,923,000
01/01/2006–31/12/2007			
Central public authority	All	Goods, Services	137,000
Local public authority	All	Goods, Services	211,000
Public entity	All	Goods, Services	422,000
Public authority/entity	All	Works, Concessions	5,278,000
01/01/2008–31/12/2009			
Central public authority	All	Goods, Services	133,000
Local public authority	All	Goods, Services	206,000
Public entity	All	Goods, Services	412,000
Public authority/entity	All	Works, Concessions	5,150,000
01/01/2010–31/12/2011			
Central public authority	All	Goods, Services	125,000
Local public authority	All	Goods, Services	193,000
Public entity	All	Goods, Services	387,000
Public authority/entity	All	Works, Concessions	4,845,000
01/01/2012–31/12/2013			
Central public authority	All	Goods, Services	130,000
Local public authority	All	Goods, Services	200,000
Public entity	All	Goods, Services	400,000
Public authority/entity	All	Works, Concessions	5,000,000
01/01/2014–31/12/2015			
Central public authority	All	Goods, Services	134,000
Local public authority	All	Goods, Services	207,000
Public entity	All	Goods, Services	414,000
Public authority/entity	All	Works, Concessions	5,186,000
01/01/2016–31/12/2017			
Central public authority	All	Goods, Services	135,000
Local public authority	All	Goods, Services	209,000
Public entity	All	Goods, Services	418,000
Public authority/entity	All	Works, Concessions	5,225,000
01/01/2018–31/12/2019			
Central public authority	Normal	Goods, Services	144,000
Central public authority	Special	Goods, Services	221,000
Local public authority	All	Goods, Services	221,000
Public entity	All	Goods, Services	443,000
Public authority/entity	All	Works, Concessions	5,548,000
01/01/2020–31/12/2021			
Central public authority	Normal	Goods, Services	139,000
Central public authority	Special	Goods, Services	214,000
Local public authority	All	Goods, Services	214,000
Public entity	All	Goods, Services	428,000
Public authority/entity	All	Works, Concessions	5,350,000

Table 32. Evolution of the European thresholds. The term *Special* refers to derogatory activity sectors.

Here are the resources used to constitute this table:

2004 http://www.marche-public.fr/Marches-publics/Textes/Directives/2004-18-CE/Montant-seuils-marches-publics.htm

2005 https://www.lemoniteur.fr/article/seuils-d-application-en-matiere-de-procedures-de-passation-des-marches-modification-des-directives-2004-17-ce-et-2004-18-ce-du-parlement-europeen-et-du-conseil.1885024

2007 https://www.lemoniteur.fr/article/seuils-d-application-en-matiere-de-procedures-de-passation-des-marches-au-1er-janvier-2006-modification-des-directives-2004-17-ce-et-2004-18-ce.729864

2009 https://www.lemoniteur.fr/article/seuils-europeens-au-1er-janvier-2008-pour-la-passation-des-marches-publics.1737704

2011 https://www.lemoniteur.fr/article/marches-publics-de-nouveaux-seuils-au-1er-janvier-2010.589449

2013 https://www.lemoniteur.fr/article/marches-publics-de-nouveaux-seuils-europeens-au-1er-janvier-2012.1050484

2015 http://www.marche-public.fr/contrats-publics/DAJ-maj-seuils-2016.htm

2017 https://www.boamp.fr/Espace-acheteurs/Actualites/Archives/Nouveaux-seuils-applicables-aux-marches-publics

2019 http://www.marche-public.fr/Marches-publics/Definitions/Entrees/Seuil.htm

2021 https://www.economie.gouv.fr/daj/marches-publics-nouveaux-seuils-europeens-applicables-au-1er-janvier-2020

C Additional TED Statistics

This section provides additional statistics describing the raw data retrieved from the TED. Some of them are presented under the form of figures in the main text.

C.1 Missing Information

Table 33 shows the number of lots with missing information, for each field in the TED, as well as the corresponding proportion. These values are shown in Figure 4 for the main fields: agent name (`CAE_NAME` and `WIN_NAME`), SIRET (`CAE_NATIONALID` and `WIN_NATIONALID`), address (`CAE_ADDRESS` and `WIN_ADDRESS`), town (`CAE_TOWN` and `WIN_TOWN`), and zipcode (`CAE_POSTAL_CODE` and `WIN_POSTAL_CODE`).

Field	Lots	%
ID_NOTICE_CAN	0	0.0
TED_NOTICE_URL	0	0.0
YEAR	0	0.0
ID_TYPE	0	0.0
DT_DISPATCH	0	0.0
XSD_VERSION	0	0.0
CANCELLED	0	0.0
CORRECTIONS	0	0.0
B_MULTIPLE_CAE	839,716	60.8
CAE_NAME	0	0.0
CAE_NATIONALID	1,154,998	83.6
CAE_ADDRESS	17,565	1.2
CAE_TOWN	0	0.0
CAE_POSTAL_CODE	18,049	1.3
CAE_GPA_ANNEX	771,943	55.8
ISO_COUNTRY_CODE	0	0.0
ISO_COUNTRY_CODE_GPA	771,943	55.8
B_MULTIPLE_COUNTRY	839,716	60.8
ISO_COUNTRY_CODE_ALL	1,380,800	99.9
CAE_TYPE	0	0.0
EU_INST_CODE	1,380,494	99.9
MAIN_ACTIVITY	125,898	9.1
B_ON_BEHALF	282,543	20.4
B_INVOLVES_JOINT_PROCUREMENT	842,210	60.9
B_AWARDED_BY_CENTRAL_BODY	842,210	60.9
TYPE_OF_CONTRACT	0	0.0
TAL_LOCATION_NUTS	933,444	67.5
B_FRA_AGREEMENT	0	0.0
FRA_ESTIMATED	1,142,255	82.7
B_FRA_CONTRACT	0	0.0
B_DYN_PURCH_SYST	873,542	63.2
CPV	71	0.01
MAIN_CPV_CODE_GPA	771,967	55.9
ID_LOT	964,768	69.8
ADDITIONAL_CPVS	643,317	46.5
B_GPA	295,339	21.3
GPA_COVERAGE	771,967	55.9
LOTS_NUMBER	4,892	0.3

Field	Lots	%
VALUE_EURO	726,472	52.6
VALUE_EURO_FIN_1	432,198	31.2
VALUE_EURO_FIN_2	432,198	31.2
B_EU_FUNDS	410,569	29.7
TOP_TYPE	191	0.0
B_ACCELERATED	1,378,410	99.8
OUT_OF_DIRECTIVES	0	0.0
CRIT_CODE	222,940	16.1
CRIT_PRICE_WEIGHT	1,022,614	74.0
CRIT_CRITERIA	324,838	23.5
CRIT_WEIGHTS	362,024	26.2
B_ELECTRONIC_AUCTION	381,550	27.6
NUMBER_AWARDS	0	0.0
ID_AWARD	38,205	2.7
ID_LOT_AWARDED	295,427	21.3
INFO_ON_NON_AWARD	1,333,106	96.5
INFO_UNPUBLISHED	0	0.0
B_AWARDED_TO_A_GROUP	944,239	68.3
WIN_NAME	114,723	8.3
WIN_NATIONALID	1,341,797	97.1
WIN_ADDRESS	380294	27.5
WIN_TOWN	240,503	17.4
WIN_POSTAL_CODE	291,095	21.0
WIN_COUNTRY_CODE	332,641	24.0
B_CONTRACTOR_SME	935,433	67.7
CONTRACT_NUMBER	558,639	40.4
TITLE	255,661	18.5
NUMBER_OFFERS	424,107	30.7
NUMBER_TENDERS_SME	1,336,939	96.8
NUMBER_TENDERS_OTHER_EU	1,361,772	98.6
NUMBER_TENDERS_NON_EU	1,363,558	98.7
NUMBER_OFFERS_ELECTR	1,261,823	91.3
AWARD_EST_VALUE_EURO	1,191,352	86.2
AWARD_VALUE_EURO	565,037	40.9
AWARD_VALUE_EURO_FIN_1	426,323	30.8
B_SUBCONTRACTED	681,245	49.3
DT_AWARD	185,687	13.4

Table 33. Completion of the TED dataset: for each field, the table indicates the number of lots without any information (column *Lots*), and the corresponding proportion (column *%*).

C.2 Number of Lots by Country

Table 34 shows the numbers of lots by country, for the 2021–2020 period. The data presented in the left-hand part of the table are used to draw Figure 1 (countries having published more than 100,000 lots), whereas the right-hand part is shown in Figure 2 (fewer than 100,000 lots).

Country	Number of lots
Poland	1,497,646
France	1,380,965
Romania	653,010
Germany	562,330
United Kingdom	432,951
Spain	346,434
Slovenia	255,036
Italy	248,259
Bulgaria	201,868
Czechia	190,679
Lithuania	184,064
Sweden	144,988
Latvia	133,171

Country	Number of lots
Belgium	95,394
Netherlands	92,712
Hungary	88,843
Finland	83,801
Denmark	81,630
Greece	66,127
Portugal	54,028
Slovakia	48,485
Norway	47,365
Austria	47,353
Estonia	44,015
Ireland	39,822
Switzerland	32,272

Table 34. Number of lots published on the TED by each country between 2010 and 2020.

C.3 Missing Identifiers

This section provides additional statistics regarding missing identifiers regarding economic agents in the TED. Table 35 focuses on buyers, Table 36 on winners, and Table 37 on both at once.

Country		2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	Overall
Austria	AT	100.00	99.91	99.33	99.46	99.37	98.46	99.47	95.97	65.81	51.38	47.57	81.98
Belgium	BE	100.00	100.00	99.90	99.02	99.32	99.11	99.66	89.13	35.35	21.89	20.49	72.09
Bulgaria	BG	100.00	98.99	77.04	10.50	4.92	0.90	0.96	0.09	0.03	0.18	0.03	12.00
Croatia	HR	100.00	100.00	0.00	0.00	0.20	0.10	0.02	0.01	0.00	0.01	0.01	0.04
Cyprus	CY	100.00	99.90	100.00	100.00	100.00	100.00	100.00	100.00	99.59	99.70	99.83	99.92
Czechia	CZ	100.00	89.17	18.05	13.67	8.05	15.87	11.85	3.83	3.82	4.47	4.39	12.98
Denmark	DK	100.00	98.68	92.87	96.04	94.25	85.50	58.64	34.90	15.60	8.93	8.90	52.86
Estonia	EE	100.00	100.00	90.42	0.37	0.38	0.54	0.07	0.10	0.19	0.26	0.05	12.03
Finland	FI	100.00	100.00	92.34	87.61	86.93	89.64	73.32	44.95	7.99	0.62	0.31	57.83
France	FR	100.00	99.97	98.96	91.75	85.24	85.65	80.79	77.05	71.50	64.02	60.30	83.64
Germany	DE	100.00	99.58	98.85	98.81	99.01	98.89	98.20	96.49	94.40	94.98	95.58	97.04
Greece	GR	100.00	98.55	86.01	97.93	98.73	97.33	94.99	91.92	90.87	89.19	86.88	92.66
Hungary	HU	100.00	100.00	100.00	2.40	0.33	0.31	1.33	0.11	0.24	0.18	0.07	20.48
Iceland	IS	100.00	99.44	100.00	97.92	100.00	100.00	64.34	77.52	74.04	47.33	2.72	60.22
Ireland	IE	100.00	100.00	98.30	99.66	99.92	93.06	92.81	78.68	86.35	70.69	70.17	88.05
Italy	IT	100.00	99.94	99.85	99.65	99.78	99.81	99.32	95.87	93.88	90.93	86.32	96.19
Latvia	LV	100.00	100.00	99.98	99.99	100.00	99.99	99.99	26.76	0.09	0.06	0.07	63.59
Liechtenstein	LI	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	78.13	75.61	76.56	90.70
Lithuania	LT	100.00	100.00	29.44	0.13	0.35	0.20	0.14	0.11	0.02	0.03	0.02	11.54
Luxembourg	LU	100.00	100.00	99.73	98.86	99.87	100.00	100.00	99.69	98.16	99.16	98.75	99.45
Macedonia	MK	100.00	100.00	90.71	100.00	95.94	98.95	95.16	92.12	86.32	56.31	75.03	81.09
Malta	MT	100.00	100.00	100.00	98.94	99.75	99.59	100.00	99.80	100.00	56.51	36.62	85.12
Netherlands	NL	100.00	99.80	98.65	95.96	27.81	10.20	11.40	8.93	11.90	7.53	8.60	37.58
Norway	NO	100.00	99.97	99.97	100.00	11.28	13.67	14.93	1.04	0.61	0.62	0.45	31.79
Poland	PL	100.00	99.28	96.80	95.97	96.85	95.42	93.88	83.37	79.04	72.71	70.84	87.78
Portugal	PT	100.00	96.46	90.04	86.26	88.75	91.78	92.65	88.01	85.11	39.41	27.31	69.23
Romania	RO	100.00	99.99	99.95	99.99	99.98	99.99	99.94	100.00	48.88	0.24	0.20	33.04
Slovakia	SK	100.00	89.50	36.52	3.25	0.24	0.25	0.19	0.00	0.02	0.04	0.00	11.95
Slovenia	SI	100.00	100.00	100.00	100.00	100.00	100.00	31.66	0.09	0.14	0.08	0.13	17.01
Spain	ES	100.00	96.71	91.06	87.94	86.02	87.03	76.30	45.72	33.72	18.86	14.55	54.18
Sweden	SE	100.00	100.00	99.65	94.29	92.66	86.05	77.38	27.20	2.87	1.19	2.47	49.38
Switzerland	CH	100.00	100.00	100.00	99.95	99.63	99.98	99.95	99.97	99.97	99.98	99.97	99.95
United King.	UK	100.00	99.97	99.78	99.88	99.29	98.68	96.74	96.55	95.41	95.88	94.87	97.80
All countries		100.00	99.39	93.71	87.22	82.45	82.55	78.45	63.73	53.18	40.81	34.56	67.48

Table 35. Proportion (%) of TED lots whose CAE identifier is missing, for each year and for the whole considered period (*Overall* column), by country and for the whole dataset (last row).

All three tables show the proportion of lots with missing identifiers, as percents, for each

country and each year of the considered period, as well as for the whole period and for all countries at once.

Country		2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	Overall
Austria	AT	100.00	100.00	100.00	100.00	100.00	100.00	99.75	93.44	94.94	64.62	59.58	88.45
Belgium	BE	100.00	100.00	100.00	100.00	100.00	100.00	98.01	95.09	93.22	92.92	93.12	96.90
Bulgaria	BG	100.00	100.00	100.00	100.00	100.00	100.00	81.84	52.39	58.35	51.80	55.13	69.80
Croatia	HR	100.00	100.00	100.00	100.00	100.00	100.00	100.00	72.70	31.34	38.31	33.69	60.02
Cyprus	CY	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	99.86	98.62	99.83	99.83
Czechia	CZ	100.00	100.00	100.00	100.00	100.00	100.00	70.42	25.68	30.57	36.02	26.45	53.86
Denmark	DK	100.00	100.00	100.00	100.00	100.00	99.98	78.47	72.75	65.26	68.34	68.09	82.65
Estonia	EE	100.00	100.00	100.00	100.00	100.00	100.00	99.97	69.16	33.49	32.45	36.38	61.07
Finland	FI	100.00	100.00	100.00	100.00	100.00	100.00	88.30	35.04	21.83	19.08	20.99	67.66
France	FR	100.00	100.00	100.00	100.00	100.00	100.00	96.92	94.18	93.75	91.50	91.80	97.16
Germany	DE	100.00	100.00	100.00	100.00	100.00	100.00	99.70	99.22	99.18	98.92	99.00	99.48
Greece	GR	100.00	100.00	100.00	100.00	100.00	100.00	99.90	98.65	97.66	93.10	93.60	97.46
Hungary	HU	100.00	100.00	100.00	100.00	100.00	100.00	99.08	99.46	92.05	41.14	33.54	82.56
Iceland	IS	100.00	100.00	100.00	100.00	100.00	100.00	99.22	90.70	81.75	62.41	16.78	69.24
Ireland	IE	100.00	100.00	100.00	100.00	100.00	100.00	74.44	46.10	53.56	51.46	41.07	74.30
Italy	IT	100.00	100.00	100.00	100.00	100.00	100.00	99.45	97.47	97.32	96.91	95.41	98.56
Latvia	LV	100.00	100.00	100.00	100.00	100.00	100.00	100.00	36.99	20.44	17.88	11.26	69.22
Liechtenstei.	LI	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	84.38	100.00	100.00	98.55
Lithuania	LT	100.00	100.00	100.00	100.00	100.00	100.00	99.96	97.88	68.01	66.28	68.01	82.20
Luxembourg	LU	100.00	100.00	100.00	100.00	100.00	100.00	98.42	91.44	95.86	97.54	91.57	97.49
Macedonia	MK	100.00	100.00	100.00	100.00	100.00	100.00	99.86	99.87	98.60	97.85	97.56	98.89
Malta	MT	100.00	100.00	100.00	100.00	100.00	100.00	100.00	99.29	51.59	40.73	41.08	72.24
Netherlands	NL	100.00	100.00	100.00	100.00	100.00	100.00	95.17	60.15	58.38	56.93	58.66	82.32
Norway	NO	100.00	100.00	100.00	100.00	100.00	100.00	100.00	54.16	54.48	55.48	59.46	80.03
Poland	PL	100.00	100.00	100.00	100.00	100.00	100.00	99.42	95.50	94.76	94.06	92.94	97.44
Portugal	PT	100.00	100.00	100.00	100.00	100.00	100.00	99.41	94.66	86.77	80.58	78.72	89.89
Romania	RO	100.00	100.00	100.00	100.00	100.00	100.00	99.98	100.00	84.22	66.31	64.56	76.83
Slovakia	SK	100.00	100.00	100.00	100.00	100.00	100.00	58.19	33.74	28.93	37.90	38.11	69.39
Slovenia	SI	100.00	100.00	100.00	100.00	100.00	100.00	42.47	14.22	27.25	31.32	34.86	39.67
Spain	ES	100.00	100.00	100.00	100.00	100.00	99.95	93.35	76.07	69.51	57.04	58.55	79.44
Sweden	SE	100.00	100.00	100.00	100.00	100.00	100.00	94.72	48.89	20.27	15.87	19.12	61.92
Switzerland	CH	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	99.95	99.98	100.00	99.99
United King.	UK	100.00	100.00	100.00	100.00	100.00	99.93	98.24	94.40	94.31	93.62	88.51	97.14
All countries		100.00	100.00	100.00	100.00	100.00	99.99	95.56	81.31	77.48	72.96	70.59	87.14

Table 36. Proportion (%) of TED lots whose winner identifier is missing, for each year and for the whole considered period (*Overall* column), by country and for the whole dataset (last row).

Country		2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	Overall
Austria	AT	100.00	99.91	99.33	99.46	99.37	98.46	99.22	90.56	65.14	50.70	46.82	81.19
Belgium	BE	100.00	100.00	99.90	99.02	99.32	99.11	97.67	84.69	31.85	19.90	18.82	70.67
Bulgaria	BG	100.00	98.99	77.04	10.50	4.92	0.90	0.96	0.08	0.03	0.17	0.03	12.00
Croatia	HR	100.00	100.00	0.00	0.00	0.20	0.10	0.02	0.00	0.00	0.01	0.01	0.04
Cyprus	CY	100.00	99.90	100.00	100.00	100.00	100.00	100.00	100.00	99.45	98.33	99.66	99.75
Czechia	CZ	100.00	89.17	18.05	13.67	8.05	15.87	10.09	2.35	1.92	3.13	2.26	11.77
Denmark	DK	100.00	98.68	92.87	96.04	94.25	85.50	50.43	29.84	9.63	5.95	6.48	50.17
Estonia	EE	100.00	100.00	90.42	0.37	0.38	0.54	0.03	0.07	0.17	0.20	0.04	12.01
Finland	FI	100.00	100.00	92.34	87.61	86.93	89.64	66.84	21.02	3.38	0.57	0.08	54.66
France	FR	100.00	99.97	98.96	91.75	85.24	85.65	79.32	74.09	68.60	59.90	55.94	82.23
Germany	DE	100.00	99.58	98.85	98.81	99.01	98.89	97.96	95.92	93.92	94.48	95.10	96.75
Greece	GR	100.00	98.55	86.01	97.93	98.73	97.33	94.89	91.22	88.84	83.66	82.32	90.72
Hungary	HU	100.00	100.00	100.00	2.40	0.33	0.31	1.31	0.03	0.15	0.10	0.02	20.44
Iceland	IS	100.00	99.44	100.00	97.92	100.00	100.00	63.57	72.09	56.84	41.76	2.36	57.45
Ireland	IE	100.00	100.00	98.30	99.66	99.92	93.06	68.84	36.50	44.85	40.55	31.71	68.92
Italy	IT	100.00	99.94	99.85	99.65	99.78	99.81	98.96	93.80	92.15	89.00	83.80	95.29
Latvia	LV	100.00	100.00	99.98	99.99	100.00	99.99	99.99	24.73	0.08	0.02	0.02	63.40
Liechtenstein	LI	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	78.13	75.61	76.56	90.70
Lithuania	LT	100.00	100.00	29.44	0.13	0.35	0.20	0.10	0.04	0.01	0.01	0.00	11.52
Luxembourg	LU	100.00	100.00	99.73	98.86	99.87	100.00	98.42	91.34	94.02	96.94	90.91	97.05
Macedonia	MK	100.00	100.00	90.71	100.00	95.94	98.95	95.02	92.08	85.49	55.52	73.58	80.55
Malta	MT	100.00	100.00	100.00	98.94	99.75	99.59	100.00	99.09	51.59	22.26	15.92	66.03
Netherlands	NL	100.00	99.80	98.65	95.96	27.81	10.20	11.28	8.47	11.39	7.21	8.26	37.40
Norway	NO	100.00	99.97	99.97	100.00	11.28	13.67	14.93	0.95	0.47	0.62	0.43	31.76
Poland	PL	100.00	99.28	96.80	95.97	96.85	95.42	93.61	81.59	77.41	71.14	69.31	87.05
Portugal	PT	100.00	96.46	90.04	86.26	88.75	91.78	92.36	85.59	77.49	37.09	25.62	67.01
Romania	RO	100.00	99.99	99.95	99.99	99.98	99.99	99.94	100.00	48.87	0.22	0.11	33.00
Slovakia	SK	100.00	89.50	36.52	3.25	0.24	0.25	0.19	0.00	0.02	0.02	0.00	11.94
Slovenia	SI	100.00	100.00	100.00	100.00	100.00	100.00	31.64	0.09	0.14	0.08	0.13	17.01
Spain	ES	100.00	96.71	91.06	87.94	86.02	87.00	74.84	43.38	30.50	13.67	10.55	51.93
Sweden	SE	100.00	100.00	99.65	94.29	92.66	86.05	77.37	27.07	2.51	0.90	2.24	49.26
Switzerland	CH	100.00	100.00	100.00	99.95	99.63	99.98	99.95	99.97	99.92	99.95	99.97	99.94
United King.	UK	100.00	99.97	99.78	99.88	99.29	98.62	95.67	92.19	90.67	90.66	85.08	95.48
All countries		100.00	99.39	93.71	87.22	82.45	82.54	77.46	61.58	51.22	38.98	32.78	66.45

Table 37. Proportion (%) of TED lots whose CAE and winner identifiers are both missing, for each year and for the whole considered period (*Overall* column), by country and for the whole dataset (last row).

D Fields of the TED dataset

This section gives the comprehensive list of all fields present in the TED CSV files used to initialize our database with CANs. We break down this list by categories, as indicated in the official TED documentation [6].

Certain fields are directly extracted from the formed filled by the CAEs, whereas others are computed based on other fields. The latter are indicated with an asterisk (*).

D.1 Notice Metadata

Table 38 presents the TED fields related to the general information of the notice.

Name	Data Type	Description	Version
ID_NOTICE_CAN	Integer	Unique ID of the contract award notice	all
TED_NOTICE_URL	String	URL of the notice on the TED Website	all
YEAR	Date	Year of publication of the notice	all
ID_TYPE	Integer	Code representing which directive type the notice falls under	all
DT_DISPATCH	Date	Date when the notice was sent to the TED for publication	all
XSD_VERSION*	R20X.SX	Version of the XML Schema definition	2.0.5
CANCELLED*	Boolean	Whether the notice was canceled (1) or not (0)	all
CORRECTIONS*	Integer	Number of later correction notices	all

Table 38. General TED fields related to the notice.

D.2 CAE Identification

Table 39 presents the TED fields focusing on the buyer(s). Some of these fields take a value among several predefined ones, which are listed below.

Name	Data Type	Description	Version
B_MULTIPLE_CAE*	Boolean	Whether the notice involves several CAEs	2.0.9
CAE_NAME	String	Name(s) of the CAE(s)	all
CAE_NATIONALID	String	National registration number(s) of the CAE(s)	all
CAE_ADDRESS	String	Postal address(es) of the CAE(s)	all
CAE_TOWN	String	City(s) of the CAE(s)	all
CAE_POSTAL_CODE	String	Zipcode(s) of the CAE(s)	all
CAE_GPA_ANNEX*	Enum	WTO Classe(s) of the CAE(s) (only for 2014–2016)	all
ISO_COUNTRY_CODE	String	ISO code for the country of the first CAE	all
ISO_COUNTRY_CODE_GPA*	String	ISO code for the *legal* country of the first CAE (only in 2014–2016)	all
B_MULTIPLE_COUNTRY*	Boolean	Whether the first CAE is related to several countries	2.0.9
ISO_COUNTRY_CODE_ALL	String	List of all other ISO country codes	2.0.9
CAE_TYPE*	Enum	Type of the contracting authority (ministry, regional, local...)	all
EU_INST_CODE	Enum	Subtype, if the CAE is an EU institution	2.0.9
MAIN_ACTIVITY	Enum	Main Activity of the CAE(s)	all

Table 39. TED fields related to the buyer.

WTO GPA Field CAE_GPA_ANNEX leverages the classification defined by the WTO Government Procurement Agreement (GPA), as detailed online[59].

CAE Type Field CAE_TYPE can contain the following values [6]:

- 1: Ministry or any other national or federal authority, including their regional of local subdivisions;
- 3: Regional or local authority;
- 4: Utilities sectors;
- 5: European Union institution/agency;
- 5A: Other international organization;

[59] http://www.wto.org/english/tratop_e/gproc_e/appendices_e.htm#ec

- `6`: Body governed by public law;
- `8`: Other;
- `N`: National or federal Agency / Office;
- `R`: Regional or local Agency / Office;
- `Z`: Not specified.

EU Institution Code If the CAE is an EU institution (CAE Type `5`), then field `EU_INST_TYPE` indicates its precise type [6]:

- `AG`: Agencies;
- `BC`: European Central Bank;
- `BI`: European Investment Bank;
- `BR`: European Bank for Reconstruction and Development;
- `CA`: European Court of Auditors;
- `CJ`: Court of Justice of the European Union;
- `CL`: Council of the European Union;
- `CR`: European Committee of the Regions;
- `EA`: European External Action Service;
- `EC`: European Commission;
- `ES`: European Economic and Social Committee;
- `FI`: European Investment Fund;
- `OB`: European Patent Office;
- `OP`: Publications office of the European Union;
- `PA`: European Parliament.

Main Activity Field `MAIN_ACTIVITY` represents the area of activity of the CAE. It relies on the Classification of the Functions of Government (COFOG)[60], which we reproduce here:

- `General public services`: Executive and legislative organs, financial and fiscal affairs, external affairs; foreign economic aid; general services; basic research; R&D related to general public services; general public services n.e.c.; public debt transactions, transfers of a general character between different levels of government.
- `Defence`: Military defence; civil defence; foreign military aid, R&D related to defence; defence n.e.c. (not elsewhere classified).
- `Public order and safety`: Police services; fire-protection services; law courts; prisons; R&D related to public order and safety; public order and safety n.e.c.
- `Economic affairs`: General economic, commercial and labour affairs; agriculture, forestry; fishing and hunting; fuel and energy; mining, manufacturing and construction; transport; communication; other industries, R&D related to economic affairs; economic affairs n.e.c.
- `Environmental protection`: Waste management; water waste management; pollution abatement; protection of biodiversity and landscape; R&D related to environmental protection.
- `Housing and community amenities`: Housing development; community development; water supply; street lighting; R&D related to housing and community amenities; housing and community amenities n.e.c.
- `Health`: Medical products, appliances and equipment; outpatient services; hospital services; public health services; R&D related to health; health n.e.c.
- `Recreation, culture and religion`: Recreational and sporting services; cultural services; broadcasting and publishing services; religious and other community services, R&D related to recreation, culture and religion; recreation; culture and religion n.e.c.
- `Education`: Pre-primary, primary, secondary and tertiary education, post-secondary non-tertiary education, education non-definable by level, subsidiary services to edu-

[60] https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Classification_of_the_functions_of_government_(COFOG)

cation, R&D; n.e.c.
- `Social protection`: Sickness and disability; old age; survivors; family and children; unemployment; housing; R&D; social protection and social exclusion n.e.c.

D.3 Notice- and Lot-Level Variables

Table 40 shows the fields describing whole contracts and lots. Fields taking values in an enumerated collection are detailed below. All monetary amounts are expressed in Euros.

Name	Data Type	Description	Version
B_ON_BEHALF	Boolean	Whether the contract involves several buyers	all
B_INVOLVES_JOINT_PROCUREMENT	Boolean	Whether the contract is a joint procurement	2.0.9
B_AWARDED_BY_CENTRAL_BODY	Boolean	Whether the CAE is a central purchasing body	2.0.9
TYPE_OF_CONTRACT	Enum	Contract related to works, supplies or services	all
TAL_LOCATION_NUTS	Enum	NUTS code for the main location of work	all
B_FRA_AGREEMENT	Boolean	Notice *declared* as related to a framework agreement (FA)	all
FRA_ESTIMATED*	Enum	Notice *estimated* as related to a FA	all
B_FRA_CONTRACT*	Boolean	Notice *estimated* as related to contracts within a FA	all
B_DYN_PURCH_SYST	Boolean	Notice involving a dynamic purchasing system	all
CPV	Enum	Main common procurement vocabulary code (2008 version)	all
MAIN_CPV_CODE_GPA*	Enum	Cleaned version of the main CPV	all
ADDITIONAL_CPVS	Enum	Additional CPV codes	all
B_GPA	Boolean	Contract covered by the Government Procurement Agreement	all
GPA_COVERAGE*	Enum	Detailed information about GPA coverage (only for 2014–2016)	all
ID_LOT	Integer	Unique identifier of the Lot	2.0.9
LOTS_NUMBER*	Integer	Number of lots in the contract (since 2009)	all
VALUE_EURO	Float	Pre-tax CAN value (€)	all
VALUE_EURO_FIN_1*	Float	Pre-tax CAN value, automatically estimated from other fields	all
VALUE_EURO_FIN_2*	Float	Pre-tax CAN value, manually estimated	all
B_EU_FUNDS	Boolean	Whether the contract is related to a project funded by the EU	all
TOP_TYPE	Enum	Type of procedure	all
B_ACCELERATED	Boolean	Whether the awarding procedure was accelerated	all
OUT_OF_DIRECTIVES	boolean	CAN published even though there was no CN	all
CRIT_CODE	Enum	Main award criterion	all
CRIT_PRICE_WEIGHT*	Float	Weight of the price criterion	2.0.9
CRIT_CRITERIA	String	Additional award criteria	all
CRIT_WEIGHTS	Float	Weights of the additional criteria	all
B_ELECTRONIC_AUCTION	Boolean	Whether an electronic auction was conducted	all
NUMBER_AWARDS*	Integer	Number of different winners for the lot	all

Table 40. TED fields related to the notices and lots.

On Behalf In field `B_ON_BEHALF`, the involvement of several buyers can be due to a joint procurement or to the buyer being a central purchasing body. This is specified in fields `B_INVOLVES_JOINT_PROCUREMENT` and `B_AWARDED_BY_CENTRAL_BODY`, respectively.

Type of Contract . Field `TYPE_OF_CONTRACT` can be one of the following:
- `W`: Works;
- `U`: Supplies;
- `S`: Services.

Main Location Field `TAL_LOCATION_NUTS` shows the main location of work, place of delivery or of performance [6]. It is a NUTS code (*Nomenclature des Unités territoriales statistiques* – Nomenclature of Territorial Units for Statistics)[61].

Relation to Framework Agreement Field `FRA_ESTIMATED` indicates the (possible) relation automatically detected between the notice and a framework agreement [6]:
- `K`: keyword "framework" found in the title or description of the notice;
- `A`: multiple awards were given per one lot;
- `C`: most of the notices which following this notice are marked as framework agreement.

[61] https://ec.europa.eu/eurostat/web/nuts/background

GPA Coverage Field `GPA_COVERAGE` indicates how the contract is cover (or not) by the Government Procurement Agreement [6]:

- `1`: covered by GPA;
- `2`: entity not covered by GPA;
- `3`: entity covered, but contract not covered by GPA;
- `4`: below-thresholds contract;
- `5`: contracting entity is not an EU public entity.

Value Fields `VALUE_EURO_FIN_1` is an estimation of the pre-tax CAN value for the case where field `VALUE_EURO` is empty. The estimation method is provided in Appendix I of [6]. Field `VALUE_EURO_FIN_2` is most often equal to `VALUE_EURO_FIN_1`, but can include an additional manual correction.

Type of Procedure Field `TOP_TYPE` shows the type of procedure used to award the contract [6]:

- `AWP`: award without prior publication of a contract notice;
- `COD`: competitive dialogue;
- `NOC/NOP`: negotiated without a call for competition;
- `NIC/NIP`: negotiated with a call for competition;
- `OPE`: open procedure;
- `RES`: restricted procedure;
- `INP`: innovative partnership.

Award Criteria Field `CRIT_CODE` indicates the criteria considered during the awarding procedure [6]:

- `L`: lowest price;
- `M`: most economically advantageous tender.

D.4 Award Metadata

Table 41 shows the fields describing awards. Fields taking values in an enumerated collection are detailed below.

Name	Type	Description	Version
`ID_AWARD`	Integer	Unique identifier for the contract award	all
`ID_LOT_AWARDED`	Integer	Unique identifier of the concerned lot	all
`INFO_ON_NON_AWARD`	Enum	Reasons why the contract was not awarded	all
`INFO_UNPUBLISHED`	Boolean	Whether some confidential information was not published	all

Table 41. TED fields related to the awards.

Contract Not Awarded Field `INFO_ON_NON_AWARD` is empty if the contract was awarded. Otherwise, it indicates why it was not awarded [6]:

- `PROCUREMENT_UNSUCCESSFUL`: no tenders or requests to participate were received, or all were rejected;
- `PROCUREMENT_DISCONTINUED`: other reasons (discontinuation of procedure).

D.5 Winning Bidder Identification

Table 42 presents the fields related to the winner(s) of the awarding process. If the contract is awarded to several winners, only the first one is supposed to be described by these fields [6].

D.6 Other CA-Level Variables

Table 43 presents the remaining fields, related to the contract award.

Name	Type	Description	Version
B_AWARDED_TO_A_GROUP	Boolean	Whether the contract was awarded to several winners	2.0.9
WIN_NAME	String	Official name of the winner	all
WIN_NATIONALID	String	National registration number of the winner	2.0.9
WIN_ADDRESS	String	Postal address of the winner	all
WIN_TOWN	String	City of the winner	all
WIN_POSTAL_CODE	String	Zipcode of the winner	all
WIN_COUNTRY_CODE	String	ISO country code of the winner	all
B_CONTRACTOR_SME	Boolean	Whether the winner is an SME	2.0.9

Table 42. TED fields related to the winner.

Name	Type	Description	Version
CONTRACT_NUMBER	Integer	Unique identifier of the contract	all
TITLE	String	Title of the contract	all
NUMBER_OFFERS	Integer	Total number of tenders received	all
NUMBER_TENDERS_SME	Integer	Number of tenders from SMEs	2.0.9
NUMBER_TENDERS_OTHER_EU	Integer	Number of tenders from other EU states	2.0.9
NUMBER_TENDERS_NON_EU	Integer	Number of tenders from non-EU states	2.0.9
NUMBER_OFFERS_ELECTR	Integer	Number of offers received electronically	all
AWARD_EST_VALUE_EURO	Float	Estimated pre-tax CA value (€)	all
AWARD_VALUE_EURO	Float	Effective pre-tax CA Value, or lowest bid (€)	all
AWARD_VALUE_EURO_FIN_1*	Float	Pre-tax CA value (€), estimated based on other fields	all
B_SUBCONTRACTED	Boolean	Whether the contract is likely to be subcontracted	all
DT_AWARD	Date	Date of contract award	all

Table 43. TED fields related to the CA-level variables.

Award Value Field AWARD_VALUE_EURO_FIN_1 is an estimation provided when AWARD_VALUE_EURO is empty. The estimation method is the same as for field VALUE_EURO_FIN_1, as described in [6].

E Additional Results

This section provides additional statistics and results related to our process described in Sections 3–5, which we propose to fix the TED errors identified in Section 2.3.

E.1 Identification Step

The following tables show the results obtained for both ground truth used during the assessment of our agent identification method in Section 4.3.

Agent Type	Full	Partial	Incorrect	None	Total
Buyer occurrences	145,726	21,107	39,682	801	207,316
	70.29%	10.18%	19.14%	0.39%	100.00%
Winner occurrences	29,668	3,571	6,651	144	40,034
	74.11%	8.92%	16.61%	0.36%	100.00%

Table 44. Results of the agent identification process for the pre-existing SIRETs, in terms of agent occurrences.

Table 44 shows the results obtained on the first ground truth, that contains only the agent occurrences whose SIRET is known in the original TED data, expressed in terms of agent occurrences. It corresponds to Figure 9a. Table 45 shows the same thing as Table 44, but in terms of unique agents. It corresponds to Figure 9b.

Agent Type	Full	Partial	Incorrect	None	Total
Unique buyers	4,070	560	1,452	33	6,115
	66.56%	9.16%	23.74%	0.54%	100.00%
Unique winners	13,187	1,638	3,295	89	18,209
	72.42%	8.99%	18.10%	0.49%	100.00%

Table 45. Results of the agent identification process for the pre-existing SIRETs, in terms of unique agents.

Table 46 shows the performance obtained for the manually constituted ground truth, previously exhibited graphically in Figure 10. As it contains only unique agents, there is no need to show the results in terms of agent occurrences, as we do for the first ground truth (pre-existing SIRETs).

Agent Type	Full	Partial	Incorrect	None	Total
Unique buyers	178	24	46	2	250
	71.20%	9.60%	18.40%	0.80%	100.00%
Unique winners	157	26	57	10	250
	62.80%	10.40%	22.80%	4.00%	100.00%

Table 46. Results of the agent identification process for the manually annotated SIRETs, in terms of unique agents (cf. Figure 10).

E.2 Clustering Step

<To be completed>

F Lexicon

In this section, we give a short definition of the main concepts related to French public procurement, TED, and more generally to the DeCoMaP project. The French translation of these expressions is given in italics.

Acceptance period / ***Période d'acceptation*** Number of calendar days (after the publication of the notice) available to the Government before for awarding a contract.

Adapted Procedure / ***Marché à procédure adaptée (MAPA)*** Procedure used to award a contract whose estimated value is below the European threshold (see also *Formalized Procedure*).

Agent Economic entity able to enter into a contract, either as a Buyer (see *Contract Authority or Entity* – CAE) or a Supplier (see winner).

Bulletin Officiel des Annonces des Marchés Publics (BOAMP) French official bulletin of public procurement notices, the national outlet used to publish French public procurement contract notices and contract award notices whose estimated value is above a certain national threshold (itself lower than the European threshold).

Call For Competition (CFC) A contract notice, a prior information notice used as a call for competition, or a qualification system with a call for competition [5].

Contract Authority or Entity (CAE) Agent acting as the buyer in a public procurement contract.

Central purchasing bodies / ***Centrale d'achat*** Contracting authority that make contracts on behalf of a CAE.

Contract Award (CA) / ***Attribution de contrat*** Result of the awarding procedure for one or several lots of the same contract. It is described in the CAN dedicated to the contract, together with the other CA of the same contract (if any).

Contract Award Notice (CAN) / ***Avis d'attribution*** Document describing the result of the awarding of a contract, i.e. the contract awards associated to this contract. It also generally contains some information regarding the contract itself, also present in the contract notice.

Common Procurement Vocabulary (CPV) / ***Vocabulaire commun pour les marchés*** European classification system aiming at describing in a normalized way the domain of the product or service that is considered in a public procurement.

Contract Notice (CN) / ***Avis de marché*** Description of a tender opportunity in the public market. Its awarding is described in a dedicated contract award notice (CAN).

Data Entry Clerk / ***Opérateur de saisie*** CAE staff in charge of entering the public procurement data into a computer system.

Dynamic Purchasing System / ***Système d'achat dynamique*** Electronic system used in public procurement, where a supplier can join any time.

Entity / ***Entité*** In the SIRENE terminology, a high level economic agent, not tied to any geographical zone, and likely to cover one or several lower level economic agents called facilities.

European threshold / ***Seuil européen*** Depending on whether its estimated value is below or above this threshold, a public procurement must follow the adapted or formalized procedures, respectively.

Facility / ***Établissement*** In the SIRENE terminology, a low level economic agent, attached to a SIRENE entity, and localized at a specific geographical point.

Formalized Procedure / ***Marché à procédure formalisée*** Procedure used to award a contract whose estimated value is *above* the European threshold (see also *Adapted Procedure*).

Framework Agreement / ***Accord cadre*** Specific type of contract between some buyers and winners, allowing to make direct purchase without competition (purchase order) or to put suppliers back to competition (subsequent markets) during a predefined period.

Government Procurement Agreement / ***Accord sur les marchés publics*** Agreement under the World Trade Organization (WTO), aiming at regulating public procurement.

Institut National de la Statistique et des Etudes Economiques (INSEE) French national institute of statistics and economic studies[24]. In charge of the SIRET and SIREN identifiers, as well as of the SIRENE database that hosts them.

Joint procurement / ***Groupement conjoint*** Combined procurement between two or more CAE or winners.

Lot Stand-alone unit of a public procurement, that is assigned separately from the other lots attached to the same contract.

Official Journal of the European Union (OJEU) / ***Journal Officiel de l'Union Européenne (JOUE)*** Outlet to publish contract notices and contract award notices of public procurement contracts with a value above the European threshold.

Open procedure / ***Procédure ouverte*** Awarding procedure allowing each supplier to submit a bid.

Public entity / ***Entité publique*** Office or department under the supervision of a local or state government.

Public procurement / ***Marché public*** Contract concluded for valuable consideration between a public or private buyer and a public or private economic operator.

Restricted procedure / ***Procédure restreinte*** Awarding procedure in which any supplier can ask to participate, but the buyer chooses who can submit an offer.

Small and Medium-sized Enterprises (SME) / ***Petites et moyennes entreprises (PME)*** Companies which employing up to 250 employees.

SIREN code / ***Code SIREN*** Unique nine digits number used to identify a company or organization in France. SIREN stands for *Système d'Identification du Répertoire des ENtreprises* (Identification system of the entity register).

SIRENE database / ***Base SIRENE (Système national d'Identification du Répertoire des ENtreprises et de leurs Etablissements)*** French database managed by the INSEE[24] (French national institute for statistics) that assigns SIRENs to entities and SIRETs to facilities.

SIRET code / ***Code SIRET*** Unique 14 digits number containing used to identify a facility in France. It contains the SIREN of the corresponding entity, followed by five digits specific to each facility attached to this entity. SIRET stands for *Système d'Identification du Répertoire des Etablissements* (Identification system of the facility register).

Tenders Electronic Daily (TED) Online version of the OJEU, dedicated to public procurement notices.

VAT Information Exchange System (VIES) Online platform providing the VAT numbers of EU companies. By extension, the European identifier itself (*Numéro de TVA intracommunautaire*).

Voluntary Ex-Ante Transparency notice (VEATs) Mandatory notice announcing that the buyer intends to place a non-competitive contract [5].

Winner / ***Gagnant*** Economic agent acting as a supplier and which was awarded a lot from a public procurement contract.