

AperTO - Archivio Istituzionale Open Access dell'Università di Torino

ITH: an open database on Italian Tenders 2016-2023

This is a pre print version of the following article:

Original Citation:

Availability:

This version is available <http://hdl.handle.net/2318/2045850> since 2025-01-11T13:43:57Z

Published version:

DOI:10.1038/s41597-024-04342-5

Terms of use:

Open Access

Anyone can freely access the full text of works made available as "Open Access". Works made available under a Creative Commons license can be used according to the terms and conditions of said license. Use of all other works requires consent of the right holder (author or publisher) if not exempted from copyright protection by the applicable law.

(Article begins on next page)

1 ITH: an open database on Italian Tenders 2016-2023

2 **Roberto Nai¹, Emilio Sulis¹, and Rosa Meo¹**

3 ¹University of Turin, Computer Science Department, Turin, 10149, IT

4 *corresponding author: Roberto Nai (roberto.nai@unito.it)

5 **ABSTRACT**

6 Governments procure large amounts of goods and services to help them implement policies and deliver public services; in Italy, this is an essential sector, corresponding to about 12% of the gross domestic product. Data are increasingly recorded in public repositories, although they are often divided into multiple sources and not immediately available for consultation. This paper describes and analyses an effort to collect and arrange a legal public administration database. The main source of interest involves the National Anti-Corruption Authority in Italy, which describes more than 3 million tenders from 2016 to 2023. To improve usability, the database is integrated with two other relevant data sources concerning information on public entities and territorial units for statistical purposes. The analysis also identifies key challenges that arise from the current Open Data catalogue, particularly in terms of data completeness. A practical application is described with an example of use. The final dataset, called Italian Tender Hub, is available in a repository with a description of its use.

7 **Background & Summary**

8 Public administration has a great impact on a Country's economy through the purchase of large quantities of goods and services
9 to implement policies and provide public services. Such expenditure is a quote of the total value of all goods and services
10 produced within a Country in a specific period, usually a year, i.e. Gross Domestic Product (GDP). In the Organisation for
11 Economic Co-operation and Development (OECD) countries, public tender expenditure as a share of GDP increased over the
12 last decades, from 11.8% of GDP in 2007 to 12.9% of GDP in 2021¹.

13 Between 2016 and 2023, Italy's public procurement expenditure hovered around 11-12% of its GDP, aligning with OECD
14 trends. This percentage fluctuated, notably during the COVID-19 pandemic, due to heightened health and infrastructure
15 demands. Italy mirrored the OECD's response, where procurement surged as governments addressed pandemic-related needs,
16 particularly in healthcare, aligning with the OECD average of 12-13%^{2,3}.

17 Comprehensive and clear knowledge of the public tender situation is of paramount importance for a national State from a
18 management perspective, increasing transparency and good governance. Moreover, better knowledge of public tender builds
19 trust in institutions, increases accountability, and leads to better quality services for citizens. Businesses can benefit by having a
20 picture of the current situation to understand possible market space. Furthermore, academic research can use such data across
21 various disciplines, from economics to management and law. These data can be very useful for training Artificial Intelligence
22 (AI) systems and for enhancing transparency through explainability techniques, as they allow AI models to learn from extensive,
23 structured information about public procurement practices, thus supporting both predictive and interpretative applications in
24 public sector analytics⁴. Finally, regulatory bodies can obtain valuable information to detect cases of corruption and fraud at an
25 early stage⁵.

26 The widespread dissemination of data through open standards have even more significant effects. A related aspect of
27 interest concerns combating possible fraud cases⁶. For instance, Italy experiences a significant quote of both frauds and court
28 inefficiencies⁷. According to the OECD⁸, Open Data (OD) can help to design better anti-corruption policies and monitor their
29 effective implementation. Increasingly, public administrations typically offer their data on institutional repositories, which can
30 be freely accessed online. Nevertheless, it is often not easy or possible to consult these databases fully. For instance, the entire
31 dataset can be divided into several web archives (web pages or websites), accessible by graphical interfaces written in different
32 web programming languages. These facilitate timely access to individual instances but do not allow overall export. In addition,
33 the datasets of interest may be in different, unconnected repositories, so the overall fruition may become somewhat difficult and
34 require attention, with possible data loss.

35 This paper aims to fill this gap by proposing a collection of data concerning a public administration at the national level,
36 focusing on a dataset of great importance such as public procurement. In our work, a large legal dataset has been collected from
37 public repositories, unified and made publicly available. The main dataset refers to public tenders collected by the national
38 body appointed for the purpose, which is the Italian AntiCorruption Authority (ANAC)⁹. The dataset has been enriched with
39 valuable information from two other relevant data sources, according to domain experts. We refer to the information on the
40 public administration of interest provided by the Database of Public Administrations (BDAP), as well as population data or the

geocode standard for referencing the administrative divisions, i.e. Nomenclature of Territorial Units for Statistics (NUTS), provided by Italian National Institute of Statistics (ISTAT). Through these additional sources, it will be possible to analyse the ANAC dataset more consistently. For instance, it is possible to investigate the territory where investments are made by including type of contracting authority that made the tender, size of the municipality where the contracting authority is located, geographical location of the investment, etc. Therefore, our effort to integrate datasets from diverse sources is motivated by the recognition that analysing public tenders data through geographical and population metrics can help to reveal critical patterns in spending and regional needs, as stated by¹⁰.

The article then describes the resulting effort to build a database called *Italian Tender Hub* (henceforth ITH) in Italy from 2016 to 2023. Following the advice of academic experts in legal studies, specialising in public tenders and procurement (hereafter referred to as *domain experts*), the time frame was selected to align closely with the validity period of the Italian public procurement code, which came into force in April 2016 and was repealed in 2023^{11,12}. Furthermore, at the suggestion of the domain experts themselves, the variables of interest of the three datasets were defined to make the study more complete and useful for those who want to use ITH extending, for example, the work proposed by¹³.

The following sections describe the methods adopted, presenting the main indicators and possibilities for use. The selection of the period also concerns reasons of both consistency, related to the abundance of complete data, and expediency, according to the legal expert's suggestions.

Methods

This Section describes the construction of the ITH database, starting with the collection of data from the institutional repositories to reconstruct a set of interest's fields. The pipeline for creating the overall database is described in Figure 1 and involves the integration of the ANAC tenders dataset and two other datasets, ISTAT and BDAP. Specifically, the three main methodological steps involve: 1) collecting the relevant data from the institutional web repositories in CSV format; 2) generating a relational database in SQL containing definitions, tables, primary keys (PK), and foreign keys (FK); 3) testing the proper functionality of the database starting from the generated SQL script; 4) publishing the complete version of the database to an easily accessible open repository. We briefly summarize the main methods adopted at each stage.

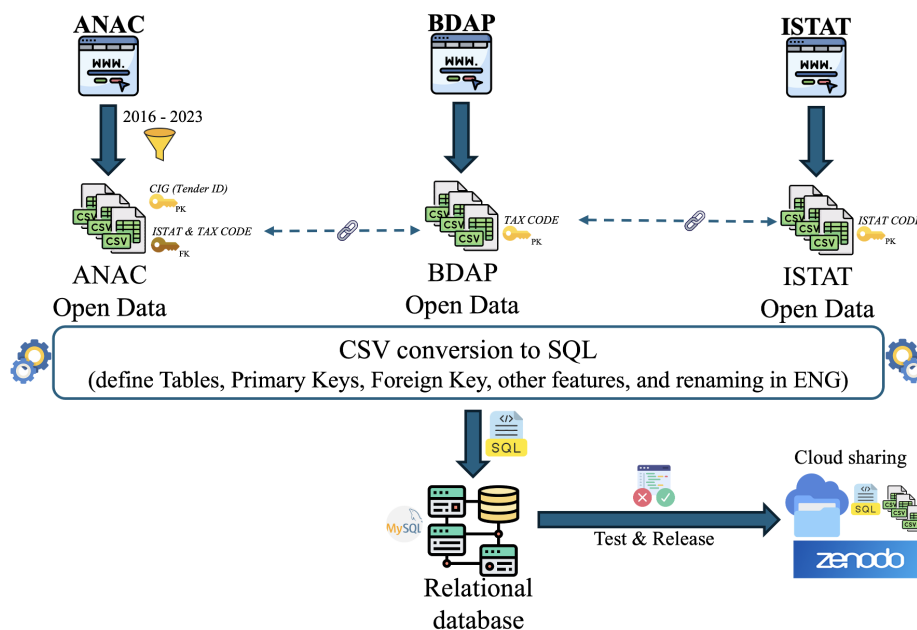


Figure 1. Workflow with data collection, database creation via SQL script, testing, and publication of the final relational database on Zenodo. Full size image available at <https://github.com/roberto-nai/ScientificData2024>

64

Data collection. The initial steps concern data collection from institutional repositories, whereas web scraping techniques have been adopted. For the ANAC website, it was necessary to develop a script for bulk downloading 222 CSV files to generate all download requests to individual files on the website. Starting with a main URL, the script generates as many dynamic URLs as there are files to be collected for each year and month of interest (in our case from 2016 to 2023). For each dynamic URL generated, an HTTPS request is executed via terminal to download the relevant file, which is then automatically unzipped by

70 the script to obtain the final CSV file on the local computer. Similarly, the same technique was adopted to collect files from the
71 BDAP and ISTAT sites.

72 **Database creation.** The data recording tool involved a relational database, which is the most diffuse type of database. The
73 organization of the data in tabular form lends itself well to manage the relationships among the datasets in our case. The
74 database can be queried using SQL query languages. Therefore, starting from the data structure of the CSV files, a script was
75 developed to convert it into an SQL file to generate the relational database and the tables with primary and foreign keys to
76 link the data together. After creating the SQL tables, we integrated ITH database using established database management
77 techniques, specifically by connecting tables based on primary and foreign keys. This standard SQL-based approach ensures
78 relational consistency and enables efficient querying across datasets¹⁴, aligning with widely recognised practices in relational
79 database design and data integration and adhering to best practices for interoperability and consistency in data management¹⁵.
80 The SQL files for generating the ITH database have been tested to be compatible with the MySQL¹⁶ versions 5.x and 8.x; we
81 opted for compatibility with MySQL which is one of the most used relational databases¹⁷.

82 **Test and release.** The SQL file containing the script to create the database and tables was tested in MySQL and, once verified
83 to be working without errors, was uploaded to the Zenodo repository¹⁸ together with the CSVs to be imported into the database
84 itself.

85 The following subsections detail the method adopted for extracting meaningful information from each data source and the
86 merging of the dataset into a single repository.

87 **Italian AntiCorruption Authority - ANAC**

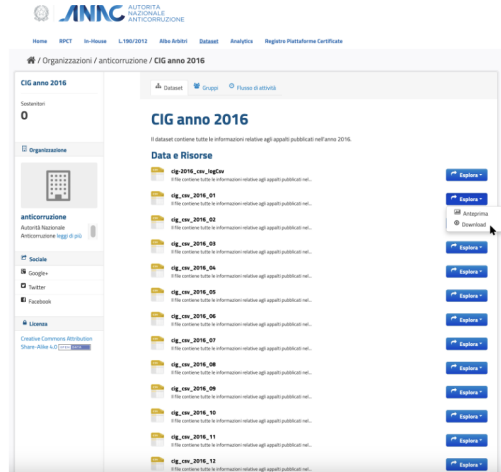
88 The main dataset concerns the government body responsible for collecting public tenders in Italy, ANAC. A specific section of
89 the ANAC website¹⁹ provides access to data in a standard view where data can be selected for categories and downloaded
90 in compressed files (Figure 2.a). In the ANAC Open Data catalogue, the main dataset is the one related to the creation of a
91 *Tender Notice*. Four other relevant datasets available include the list of the *Contracting Authorities* (CA) that have created
92 a tender, the list of tenders that received an *Award*, the *Economic Operators* (EO) that awarded a tender, and the *activities*
93 related to a tender after the awarding process (e.g. *contract-start*, *contract-end*, *subcontract*, etc.). For each of these four other
94 datasets, we provide a brief description. CAs are public bodies or entities that act on behalf of one or more public bodies
95 and are responsible for acquiring goods, services or works through tendering procedures. These authorities are in charge of
96 managing and supervising the entire public tender process. CAs can be of three types: Central (e.g. ministries), Regional and
97 Local (e.g. municipalities), and other entities (e.g. hospitals). Awards contain the list of awarded tenders with the final amount
98 awarded, the date of the award and the EOs that was awarded the tender. EOs can be sole enterprises, artisans, partnerships or
99 capital companies, cooperatives, etc. Each tender and its related activities are identified by a 10-character alphanumeric feature
100 called *CIG*. CAs and EOs are identified by their *tax code* (an alphanumeric string); CAs also have a unique *ISTAT code*.
101 Figure 2.b details an excerpt of *Tender Notice* and *Award* where not all tenders are awarded to an EO (cells *DATE_AWARD* and
102 *TENDER_AWARD* empty).

103 **Italian National Institute of Statistics - ISTAT**

104 A relevant aspect concerns identifying the scope of each tender's administrative aggregation. The National Institute of Statistics
105 (ISTAT)²⁰ provides a wide range of statistical information concerning Italy; among them, it's possible to find the *Nomenclature*
106 *of Territorial Units for Statistics* (NUTS)²¹ and the distribution of *population per municipality*. The NUTS system is organized
107 into three hierarchical levels: NUTS 1 includes socio-economic regions, such as large economic areas; NUTS 2 concerns
108 a smaller region for applying regional policies, like provinces or large metropolitan areas; NUTS 3 involves the smallest
109 areas, such as regions, provinces and municipalities²². The distribution of inhabitants can be of interest for understanding, for
110 example, the quote of investment per population in a certain area²³. The inclusion of NUTS and population in ITH facilitates,
111 for instance, the comparison and analysis of territorial investments at region/province/municipality level, whereas NUTS and
112 population are identified by the *ISTAT code* (an alphanumeric string) of the corresponding municipality.

113 **Database of Public Administrations - BDAP**

114 The Database of Public Administrations (BDAP)²⁴ includes relevant information on several aspects of public administrations,
115 such as the organizational structure (e.g. municipality, school/university, hospital, etc.), geographical coverage (North, Central
116 or South Italy), and budgetary information. Such information aims to enhance transparency, improve operational efficiency, and
117 support policy development. Including BDAP in ITH facilitates, for instance, comparison and analysis of investments by type
118 of authority; categorizations described above are identified by the *tax code* (an alphanumeric string) of the municipality they
119 refer to.



(a)

CIG	OBJECT	CONTRACTING AUTHORITY	DATE PUBLICATION	DATE AWARD	TENDER AMOUNT	TENDER AWARD	CITY	TENDER TYPE
5686540940	ATTREZZATURE SERVIZIO ANESTESIA E RIANIMAZIONE	AZIENDA UNITA' SANITARIA LOCALE N. 6 SANLURI	13/03/2014	27/03/2014	€48.980,00	€48.980,00	N/A	FORNITURE
997686731A	P23-216-TLP-LATRONICO-PZ-RIPRISTINO STATICO FUNZIONALE DEL SOLAIO DI COPERTURA	3 REPARTO GENIO A.M.BARI	26/07/2023	N/A	€87.397,16	N/A	Bari	LAVORI
A040977869	POLIZZA ASSICURATIVA ALL RISKS BENI IMMOBILI E MOBILI 24 MESI + 12 MESI RINNOVO	COMUNE DI LONGARE	19/12/2023	N/A	€60.357,72	N/A	Vicenza	SERVIZI
A03FEA72CE	CONVENZIONE PER LA FORNITURA MATERIALI DI PULIZIA, MONDOUSE ED IGIENE PERSONALE PER I COMANDI/ENTI DELLA GIURISDIZIONE DEL COMANDO INTERREGIONALE MARITTIMO SUD E DELLE UO NN. DIPENDENTI DALL'AREA OPERATIVA DI CINCPAN	DIREZIONE DI COMMISSARIATO MARINA MILITARE - TARANTO	19/12/2023	N/A	€215.000,00	N/A	Taranto	FORNITURE
0075606809	SS.SS. NN.33-336-341 E N.S.A. N.26	ANAS - SOCIETA' PER AZIONI	17/09/2007	N/A	€50.000,00	N/A	Roma	LAVORI
99124035B9	MESSA IN SICUREZZA DELL'AUDITORIUM SCUOLA 'ALDO MORO' DI VIA XXIV MAGGIO - OPERE EDILI	COMUNE DI CISLAGO	23/06/2023	18/07/2023	€129.089,87	€119.690,12	Varese	LAVORI
6861416B98	S.C. 20016693 CAVO 3X1X185 MM ² AREA IPIHDEX ISOLATO IN XLPE (D HPT) 12/20 KV. TIPO ASPIRALE VISIBILE, CONDUTTORI E SCHERMO IN ALLUMINIO, ISOLAMENTO ESTRUSO A SPESSORE RIDOTTO, GUAINA TERMOPLASTICA RESISTENTE ALL'URTO E ALL'ABBRASIONE, ALTRE CARATTERISTICHE SECONDO S.T. ACEA DISTRIBUZIONE DMX7. ED. LUGLIO 2016 (IMPORTO COMPRENSIVO DELL'OPZIONE DI ESTENSIONE DEL 20%)	ARETI S.P.A.	13/12/2016	23/12/2016	€285.465,60	€229.488,00	Roma	FORNITURE

(b)

Figure 2. (a) A page of the ANAC website for downloading CSV files (by year and month); (b) CSV file data preview. Full size image available at <https://github.com/roberto-nai/ScientificData2024>

120 Data Records

121 This section describes the main ITH database features based on the processing steps described in the previous section. The
 122 complete dataset containing the data described and analysed in this paper is publicly available as a Zenodo repository¹⁸. Figure 3
 123 provides a general overview of the database tables, with primary and foreign keys, categorised in four *sections* according to
 124 their contents, i.e. *Main*, *Registries*, *Award*, and *Activities after award*.

125 The *Main* tables are TENDER_NOTICE (Figure 3.a) which contains basic data on tenders and its related tables PUBLICATIONS,
 126 WORK_CATEGORY and PNNR_REWARD_MEASURES, all linked by the primary/foreign key CIG; the table is linked to the
 127 registry tables (Figure 3.b) via various foreign keys (e.g., CONTRACTING_AUTHORITIES with *tax code* of the CA), that
 128 complement it.

129 The *Registries* tables (Figure 3.b) contain precisely the registries of the CAs, linked via the foreign key *istat code* and *tax*
 130 *code* to the ISTAT and BDAP tables; this section also contains the tables with the codes useful to classify a tender (EO selection
 131 codes and tender execution codes).

132 The *Award* tables contain data on tenders awarded by an EO: AWARDS, ECONOMIC_OPERATOR and CERTIFICATES_SOA,
 133 linked to the tender table via foreign key CIG (see TENDER_NOTICE).

134 The *Activities after the award* tables (CONTRACT_START, CONTRACT_END, VARIANTS, etc.) contain data on activities
 135 carried out through the tender after the award phase; these tables are linked to the AWARDS table via foreign keys CIG and
 136 ID_AWARD.

137 We provide an idea of the main tables' dimensions. In particular, table TENDER_NOTICE contains 3,336,360 tenders,
 138 CONTRACTING_AUTHORITIES contains 435, 18 distinct CAs that created the tender notice, AWARD contains 1,830,388
 139 awarded tenders, and ECONOMIC_OPERATOR contains 1,828,831 distinct EOs that awarded the tenders.

140 In the following, we describe each of the 22 tables included in ITH (in alphabetical order):

- 141 1. AWARDS: Data on the awarding of a tender. For each tender, it's possible to have multiple awards identified by a different award
 142 identifier (*id_award*); multiple awards occur in the event of the early ending of an award's revocation.
- 143 2. CERTIFICATES_SOA: Data on the SOA attestation, i.e. a document issued by a Certification Company following an investigation in
 144 which the possession of the requirements based on work carried out in the previous period. The certificate serves the company to

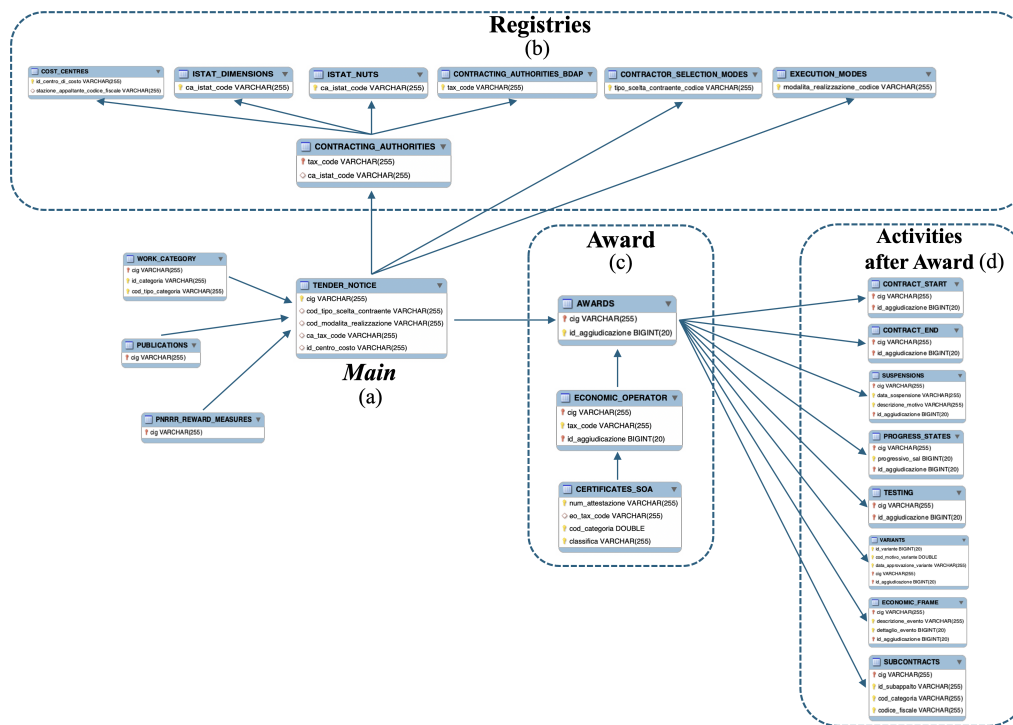


Figure 3. Data model of the ITH database. Full size image available at <https://github.com/roberto-nai/ScientificData2024>

- 145 prove, during the tender, its capacity to perform works belonging to a certain category of work and up to a certain amount.
- 146 3. **CONTRACT_END**: Data on the end of the contract between CA and EO referring to a specific tender;
- 147 4. **CONTRACT_START**: Data on the start of the contract between CA and EO referring to a specific tender;
- 148 5. **CONTRACTING_AUTHORITIES**: List of CA who created the tender; the two main primary keys (*tax_code* and *istat_code*) are used
- 149 to connect with **BDAP** and **ISTAT**;
- 150 6. **CONTRACTING_AUTHORITIES_BDAP**: **BDAP** data organizational structure of CA and geographical coverage, identified by the
- 151 primary key *tax_code*;
- 152 7. **CONTRACTOR_SELECTION_MODES**: List of criterion code - description for awarding a tender to an EO (e.g., classic tender, low
- 153 budget, multi-year agreement, etc.);
- 154 8. **COST_CENTRES**: List of cost centre code - description on which a CA associates tender costs (e.g., public transport, local police,
- 155 green);
- 156 9. **ECONOMIC_FRAME** List of the final costs of each tender (e.g. advertisement, consulting, material purchase, etc.);
- 157 10. **ECONOMIC_OPERATORS**: List of EOs who awarded tenders over time, after some verifications **CERTIFICATES_SOA**, identified
- 158 by the primary key *tax_code*;
- 159 11. **EXECUTION_MODES**: Cost centre codes - description on which a CA associates tender costs;
- 160 12. **ISTAT_DIMENSIONS**: Population data by municipality identified via *istat_code*;
- 161 13. **ISTAT_NUTS**: List of NUTS data by municipality identified via *istat_code*;
- 162 14. **PNRR_REWARD_MEASURES**: List of measure code - description listing the extra scoring criteria that can be attributed in the
- 163 **PNRR**²⁵ tender awarding rankings (e.g. recruitment of staff with special needs, number of gender-equal employees, etc.);
- 164 15. **PROGRESS_STATES**: Data on the progress (intermediate steps) of a tender;
- 165 16. **PUBLICATIONS**: Date of publication of the tender notice on official CA communication channels (e.g. **GURI**²⁶, **TED**²⁷, etc.);
- 166 17. **SUBCONTRACTS**: Data on eventual subcontracting between an EO and other suppliers to realize a part of the tender;

- 167 18. **SUSPENSIONS**: Data on eventual suspension of work on a tender (e.g.: weather problems, project problems, etc.);
- 168 19. **TENDER_NOTICE**: *Main* data, list of the tender created by CA in the various Italian regions from 2016 to 2023; starting from this
169 table, most of the other tables are linked via the primary key *CIG*.
- 170 20. **TESTING**: Data about final checks of the Work, Supplies or Services of a tender;
- 171 21. **VARIANTS**: Data on variants (changes) to the originally awarded tender (e.g.: variants for urgent project requirements);
- 172 22. **WORK_CATEGORY**: List of category codes - description with the categorisation of each tender (e.g.: motorway work, bridge, viaduct,
173 etc.).

174 Main data overview

175 This section presents a short description of the most relevant features from three tables of ITH, whereas their features are listed
176 in Table 1.

177 In **TENDER_NOTICE** table, each tender is identified by an alphanumeric value called *CIG* (the key ID value), used to
178 connect most of the remaining tables. The main distinction between tenders is their *type* and *sector*: types can be “Services”
179 (S), “Supplies” (U) or “Works” (W) while sectors can be “Ordinary” (O) or “Extraordinary” (E) based on whether they are
180 planned or due to extraordinary events (e.g. floods, earthquakes, etc.). All the types of tender are described by the CPV code,
181 i.e. *Common Procurement Vocabulary*²⁸. These categories are organized into an ontology (in a hierarchical organization) whose
182 elements are identified by codes; using the first two digits of the codes (that correspond to the upper part of the ontology and
183 the coarsest grain categories) they provide the CPV *divisions* useful to distinguish the product categories purchased by CAs
184 (e.g. “90” represents cleaning services while “9040” sewer cleaning). A tender can be defined inside a *framework agreement*,
185 meaning that the CA and EO have a previous agreement to provide services for further tenders for a defined duration of time
186 (e.g., 1 - 5 years). Often, a tender is split into *lots*, with a lower amount. Finally, each tender has a well-defined *selection*
187 *criterion* to choose the EO who will be awarded, and *implementation criterion* which the winning EO will have to comply with.

188 The table **AWARDS** contains the list of relevant features related to the tender award, with the awarding entity, date, amount,
189 etc. As expected, non-awarded tenders are not reported in this table (so they are only available in the **TENDER_NOTICE** table).

190 The table **CONTRACTING_AUTHORITIES** includes information about the name of the CA as well as the main keys
191 to link to the other tables. In this respect, the *tax code* is used to know the type of CA by joining this table with
192 **CONTRACTING_AUTHORITIES_BDAP** table from Registries, while *ISTAT code* is used to join this table with
193 **ISTAT_DIMENSIONS** and **ISTAT_NUTS** to investigate the geographical dimensions.

194 Technical Validation

195 This section presents a validation of ITH to support the technical quality of the dataset and some practical applications with
196 usage examples. First, we verified that the dataset appears to be complete in essential parts, such as the fields in Table 1, as
197 these are mandatory in the submission form compiled by the CAs. It is important to note that CAs complete standardised
198 forms in which values are either preset or entered manually, ensuring that fields left empty are expected, and populated fields
199 conform to the required values. Thus, while missing value checks are fundamental, they are sufficient in this context due to the
200 structured and consistent nature of the input forms.

201 The technical validation of our dataset primarily involved ensuring data completeness by identifying and addressing missing
202 values, a foundational approach widely adopted in data quality assessments^{29,30}. This check, a standard practice in database
203 validation, enables the identification of gaps that could affect data reliability. Upon closer examination, we identified two types
204 of missing data: (i) fields where missing values are expected, as specific entries are not required, and (ii) fields with missing
205 values likely due to data entry gaps, a common issue in data quality management. In the following, we focus on both cases.

206 i) *Data not expected*. An in-depth analysis with domain experts established that it is normal for several fields to be empty.
207 For instance, some entries (e.g., the code for PNRR) were added later, and previous notices could not contain this information.
208 In fact, we notice in the main table **TENDER_NOTICE**, features with missing value are the flag PNRR tender (76%). Discussion
209 with domain experts allows us to check how it is correct that these values are missing, since PNRR started only in 2022.

210 Regarding the table **AWARDS**, the two features with high missing value is correct they are empty. For instance, the minimum
211 and maximum values of the discount offered with respect to the tender amount (55%) and the award criterion; following domain
212 experts, the missing discount can be considered as not applied.

213 ii) *Omission in data entry*. In other minority cases, the missing field is probably due to an omission by those who were
214 supposed to add the information to the database. For instance, the ISTAT code for geographic aggregation in the table
215 **CONTRACTING_AUTHORITIES** has 4% missing fields. We notice how these values can be reconstructed by linking other
216 fields in the same dataset (i.e., via the location name, instead of the corresponding PK).

Table	Feature	Description
T_N	CIG	PK: alphanumeric value
	Tender object	Textual summary of the tender
	Framework agreement between PA and EO	1 if yes, else 0
	Number of lots	Integer value {1..n}
	Tender type	Supplies (U) Works (W) Services (S)
	Tender area	Ordinary (O) Special (S)
	Tender amount	Float value
	Date of publication	Date in format yyyy-mm-dd
	EO selection mode	Integer value {1..122}
	Execution mode	Integer value {1..19}
	Region ²¹	Italian region names + Central Government
	CPV ²⁸	String ID (XX000000-Y)
	CPV division code (first two digits of CPV) ²⁸	String ID (XX)
PNNR flag	1 if yes, else 0	
AW	<i>CIG_{FK}</i> + AWARD_ID	PK: alphanumeric value
	EO consortium (group of EOs)	1 if it's a group of EOs, else 0 (individual)
	Award date	Date in format yyyy-mm-dd
	Awarded amount (bid amount)	Float value
	Awarded amount drop (bid drop)	Float value
	Number of bids admitted	Integer value {1..n}
Subcontracting admitted	1 if yes, else 0	
C_A	Tax Code	PK: alphanumeric value (reference to BDAP.Tax Code)
	ISTAT Code	Alphanumeric value (reference to ISTAT.Code)
	CA denomination	Textual string (e.g. "Municipality of Vinovo")
BDAP	CA Tax Code	PK: alphanumeric value
	Region	Textual string (e.g. "Piedmont")
	Province	Textual string (e.g. "Turin")
	Zone	Textual string (e.g. "North", "South", "West", "East")
	ISTAT Code	Alphanumeric value (reference to ISTAT.Code)
ISTAT	CA Type	Textual string (e.g. "Municipality", "University", "Hospital")
	Istat Code	PK: alphanumeric value
	NUTS Code	Alphanumeric value (e.g. ITC11 for "Turin" province)
	Surface area (sq. km)	Float value
	Resident population (at 31 December 2023)	Integer value
	Littoral zone	1 if yes, else 0
Island zone	1 if yes, else 0	

Table 1. Main features of the tables TENDER_NOTICE (T_N), AWARD (AW), CONTRACTING_AUTHORITIES (C_A) from ANAC, and the integrated datasets BDAP and ISTAT.

217 Usage Notes

218 Load the database

219 The ITH database content is easily accessible through the Zenodo repository¹⁸ that contains an SQL schema of the database, as
220 well as a CSV version of each table constituting this database. As a first step, execute the SQL script `ITH_db_catalogue.sql`
221 in MySQL to create the database and its tables with primary and foreign keys. Next, import the CSV files into the various tables
222 via the MySQL functionality (the names of the tables to be populated correspond with the names of the CSV files). Figure 4

223 represents the command to be executed in MySQL to import each CSV file into the corresponding database table. Note the
224 following order of data import (to avoid errors between primary and foreign keys): CONTRACTING_AUTHORITIES_BDAP,
225 ISTAT_DIMENSIONS, CONTRACTING_AUTHORITIES, CONTRACTOR_SELECTION_MODES, EXECUTION_MODES,
226 COST_CENTRES, TENDER_NOTICE, all other tables.

227 Please note that some CSV files (e.g.: TENDER_NOTICE) are about 4.5 GB in size, so depending on the configuration of
228 the workstation, importing data into the database may take several minutes, and it could be necessary to set some parameter
such as execution timeout, importable file size, etc.

```
LOAD DATA INFILE 'TENDER_NOTICE.csv'  
INTO TABLE TENDER_NOTICE  
FIELDS TERMINATED BY ';'   
ENCLOSED BY '"'   
LINES TERMINATED BY '\n'  
IGNORE 1 LINES;
```

Figure 4. Example of an SQL query to import a CSV file into the corresponding database table; the text in blue is related to the standard SQL syntax while the text in orange (between the quotation marks) and black (without quotation marks) is the customisation of the query parameters. Full size image available at <https://github.com/roberto-nai/ScientificData2024>

229

230 Queries on the database

231 The example in Figure 5 refers to an example of a SQL query that extracts the tenders of the year 2016 by merging them with CA,
232 award data, EO who awarded the contract, ISTAT and BDAP data on the region/province/municipality of the CA. The query se-
233 lects all columns from the tables TENDER_NOTICE, CONTRACTING_AUTHORITIES, AWARDS, ECONOMIC_OPERATOR,
234 ISTAT_NUTS, ISTAT_DIMENSIONS, and CONTRACTING_AUTHORITIES_BDAP. It performs a series of *left joins* to
combine these tables based on matching primary/foreign keys.

```
SELECT  
TENDER_NOTICE.*, CONTRACTING_AUTHORITIES.*, AWARDS.*, ECONOMIC_OPERATOR.*,  
ISTAT_NUTS.*, ISTAT_DIMENSIONS.*, CONTRACTING_AUTHORITIES_BDAP.*  
FROM  
TENDER_NOTICE  
LEFT JOIN  
CONTRACTING_AUTHORITIES ON TENDER_NOTICE.ca_tax_code = CONTRACTING_AUTHORITIES.tax_code  
LEFT JOIN  
AWARDS ON TENDER_NOTICE.cig = AWARDS.cig  
LEFT JOIN  
ECONOMIC_OPERATOR ON TENDER_NOTICE.cig = ECONOMIC_OPERATOR.cig  
LEFT JOIN  
ISTAT_NUTS ON CONTRACTING_AUTHORITIES.ca_istat_code = ISTAT_NUTS.ca_istat_code  
LEFT JOIN  
ISTAT_DIMENSIONS ON CONTRACTING_AUTHORITIES.ca_istat_code = ISTAT_DIMENSIONS.ca_istat_code  
LEFT JOIN  
CONTRACTING_AUTHORITIES_BDAP ON CONTRACTING_AUTHORITIES.tax_code = CONTRACTING_AUTHORITIES_BDAP.tax_code  
WHERE  
TENDER_NOTICE.anno_publicazione = 2016;
```

Figure 5. Example of an SQL query to extract the tenders of the year 2016 by merging various tables; the text in blue is related to the standard SQL syntax while the text in black is the customisation of the query parameters. Full size image available at <https://github.com/roberto-nai/ScientificData2024>

235

236 Practical use of the dataset

237 To describe a possible analysis from the dataset, we provide an example of use: the analysis of public tenders for security
238 services (i.e., CPV division number 79) for CA types. First, the TENDER_NOTICE table has been filtered for *cpv_division*
239 value “79”; then, the CONTRACTING_AUTHORITIES table was cross-referenced with ISTAT tables (via *istat_code*) and
240 BDAP table (via *tax_code*) to obtain information on the type and geographical location of the CAs themselves. Some summary
241 statistics have been extracted and discussed with domain experts to demonstrate the usefulness of ITH. Figure 6.a illustrates the

242 distribution of the 5 most present NUTS, Figure 6.b illustrates the aggregation of NUTS by area, and Figure 6.c illustrates
 243 the distribution of CA types. This result has been possible thanks to integrating the ANAC dataset with the ISTAT and
 244 BDAP datasets in ITH. The script that extracts other statistics and aggregations to the final database is publicly available
 245 (<https://github.com/roberto-nai/ANAC-OD-CASESTUDY>). A more in-depth work on these data is detailed in^{31,32}.

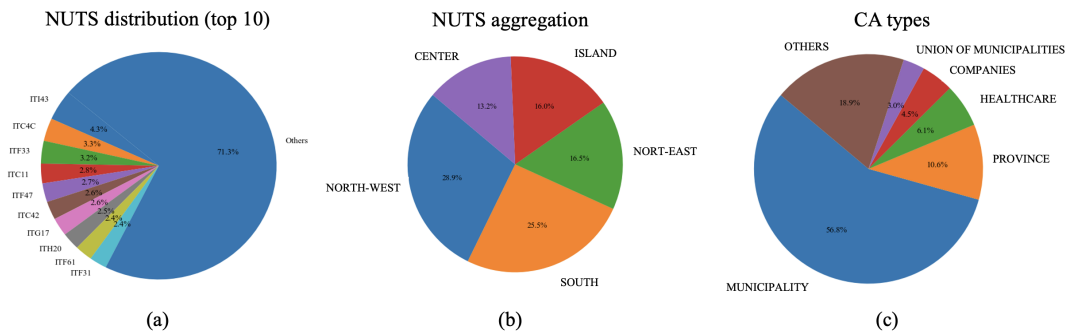


Figure 6. Distribution of NUTS (geographical location of CAs) and CA types; (a) distribution of the 5 most present NUTS; (b) aggregation of NUTS by area; (c) distribution of the 5 most present CA types. Full size image available at <https://github.com/roberto-nai/ScientificData2024>

246 To further describe the potential of ITH, we propose an analysis with georeferencing of tenders, the distribution of municipal
 247 expenditure by region and inhabitants. By cross-referencing the data with ISTAT and BDAP tables, it was possible to identify
 248 the expenditure of small, medium and large municipalities for every Italian region. As a result, it was possible to generate a
 249 map of Italy with darker colours for regions that invested more and lighter colours for regions that spent less, as described in
 250 Figure 7.

251 Extensions of use

252 Expanding the database with additional data sources can enhance analytical capabilities and offer deeper insights into socio-
 253 economic and judicial contexts. However, our examination found that the sources considered contain datasets that are currently
 254 not fully usable or directly manageable. For instance, we highlight two potentially promising sources: the first is *OpenCoesione*
 255 (<https://opencoesione.gov.it/en>), which provides data on projects funded by cohesion policies in Italy, particularly those
 256 supported by EU and national funds to reduce socio-economic and territorial disparities. This open government initiative aims
 257 to include detailed information on project goals, implementing bodies, intervention areas, and project progress. The second
 258 source is the *Administrative Justice* web platform (<https://www.giustizia-amministrativa.it>), which includes complaints and
 259 verdicts from Regional Administrative Courts (TAR) against Public Administrations, particularly on public tenders. This
 260 dataset includes documents on disputes related to procurement and awarding procedures, offering valuable legal insights into
 261 administrative conflicts. Unfortunately, it currently cannot be automatically linked to ITH, as there are no identifiers to connect
 262 a judgment to the related tender, given the data is unstructured text, as explored in^{31,33}.

263 Code availability

264 The scripts for downloading, analysing and merging ANAC's Open Data catalogue with ISTAT and BDAP were developed
 265 in Python 3.12 and are publicly available. The script for downloading the various datasets in CSV format and merge them
 266 (<https://github.com/roberto-nai/ANAC-OD-DOWNLOADER>) uses the external libraries *requests* (<https://pypi.org/project/requests>)
 267 and *urllib3* (<https://pypi.org/project/urllib3>) to automate requests via HTTPS while the script for data analysis and visualisation
 268 uses the external libraries *pandas* (<https://pypi.org/project/pandas>) and *matplotlib* (<https://pypi.org/project/matplotlib>). The
 269 script to analyse and transform the dataset from CSV to SQL (<https://github.com/roberto-nai/ANAC-OD-ANALYSER>) uses the
 270 external libraries *pandas* and *Openpolis* (<https://github.com/openpolis>) to generate the Italian map. The MySQL client used
 271 to create and populate the database is the freeware tool *MySQL Workbench* (<https://dev.mysql.com/downloads/workbench>).
 272 The script for the data analysis described in the section "Practical use of dataset" uses the aforementioned external libraries
 273 *matplotlib* and *pandas*. The complete dataset containing the data described and analysed in this paper is publicly available as a
 274 Zenodo repository¹⁸.

275 References

276 1. OECD. *Government at a Glance 2023* (OECD Publishing, 2023).

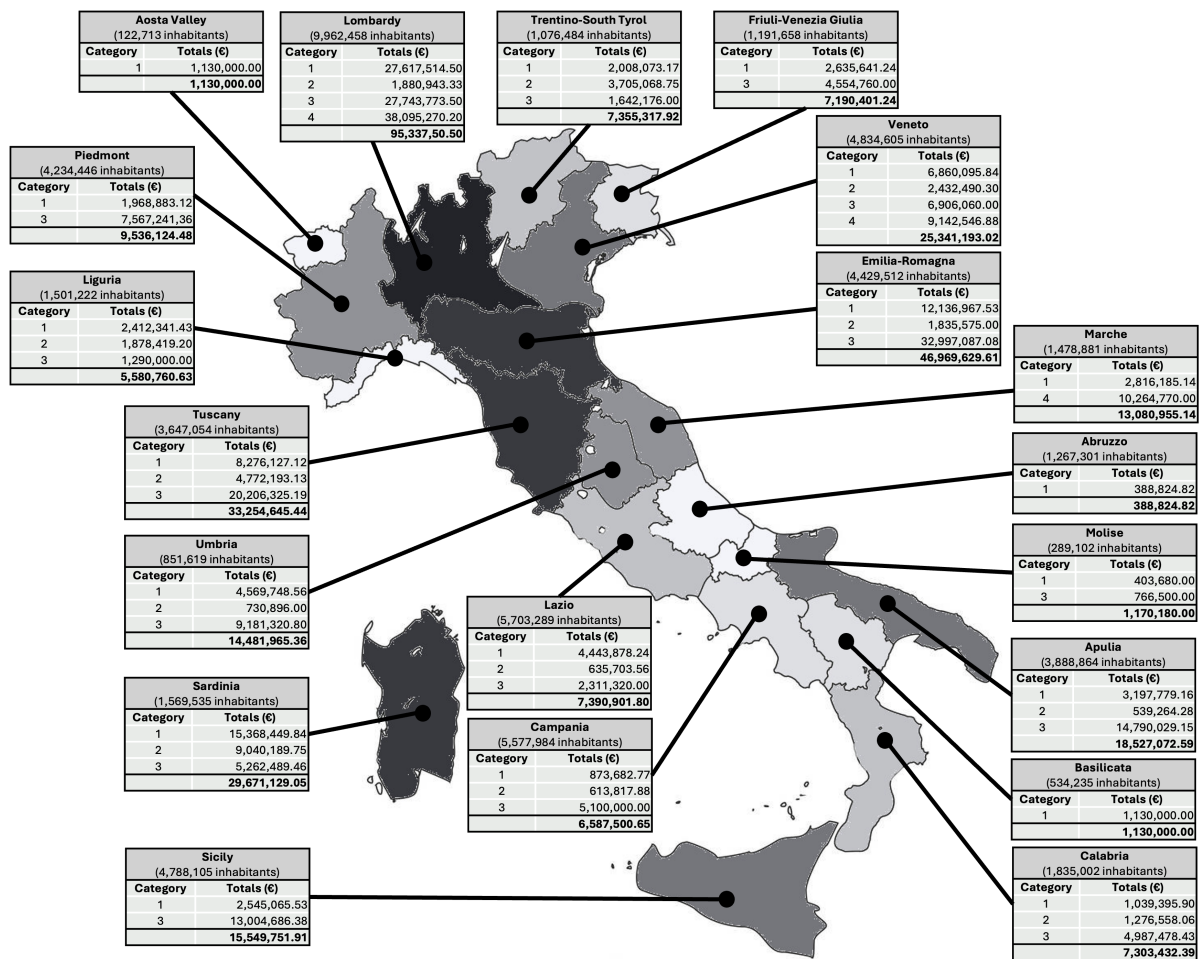


Figure 7. Map of Italy with darker colours for regions with higher investments and lighter colours for regions with lower expenditures, by CA category (sourced from BDAP). Full size image available at <https://github.com/roberto-nai/ScientificData2024>

277 2. OECD. *Government at a Glance 2021* (OECD Publishing, 2021).

278 3. OECD. Country statistical profile: Italy 2023 (2023).

279 4. Meo, R., Nai, R. & Sulis, E. Explainable, interpretable, trustworthy, responsible, ethical, fair, verifiable AI... what's next?
 280 In Chiusano, S., Cerquitelli, T. & Wrembel, R. (eds.) *Advances in Databases and Information Systems - 26th European*
 281 *Conference, ADBIS 2022, Turin, Italy, September 5-8, 2022, Proceedings*, vol. 13389 of *Lecture Notes in Computer Science*,
 282 25–34, 10.1007/978-3-031-15740-0_3 (Springer, 2022).

283 5. Nai, R., Sulis, E. & Meo, R. Public procurement fraud detection and artificial intelligence techniques: a literature review.
 284 In *Companion Proceedings of the 23rd International Conference on Knowledge Engineering and Knowledge Management*,
 285 *Bozen-Bolzano, Italy, September 26-29, 2022*, 45–56 (2022).

286 6. Passas, N. Corruption in the procurement process/outourcing government functions: Issues, case studies, implications.
 287 *Rep. to Inst. for Fraud Prev. Boston: Northeast. Univ.* (2007).

288 7. Decarolis, F., Mattera, G. & Menon, C. Do local court inefficiencies delay public works? <https://doi.org/https://doi.org/10.1787/fe4dd331-en> (2023).

290 8. Marczynski, A. & Marín, J. M. *Compendium of good practices on anti-corruption for OGP action plans* (Transparency
 291 International, 2018).

292 9. Autorità Nazionale Anti Corruzione - Institutional website. ANAC. <https://www.anticorruzione.it/en/mission-e-competenze>.

293

- 294 **10.** Pereira, R. & Furtado, F. Scaling behaviour of public procurement activity. *PLOS ONE* **14**, e0225536, [10.1371/journal.pone.0225536](https://doi.org/10.1371/journal.pone.0225536) (2019).
- 295
- 296 **11.** Government, I. Legislative decree n. 50 of 2016 (public procurement code). Italian Government (2016). Subsequently
297 replaced by Legislative Decree n. 36 of 2023.
- 298 **12.** Government, I. Legislative decree n. 36 of 2023 (2023).
- 299 **13.** Decarolis, F. & Giorgiantonio, C. Corruption red flags in public procurement: new evidence from italian calls for tenders.
300 *EPJ Data Sci.* **11**, 16, [10.1140/epjds/s13688-022-00316-1](https://doi.org/10.1140/epjds/s13688-022-00316-1) (2022).
- 301 **14.** Silberschatz, A., Korth, H. F. & Sudarshan, S. *Database System Concepts* (McGraw-Hill, 2011).
- 302 **15.** Doan, A., Halevy, A. Y. & Ives, Z. G. *Principles of Data Integration* (Morgan Kaufmann, 2012).
- 303 **16.** Oracle Corporation. *MySQL: The World's Most Popular Open Source Database*. Oracle Corporation (2023). Version 8.0.
- 304 **17.** DB-Engines. DB-Engines ranking of relational DBMS (2024). Accessed: 2024-06-20.
- 305 **18.** Nai, R. & Sulis, E. ITH: An open database on Italian tenders 2016 - 2023, [10.5281/zenodo.12179651](https://doi.org/10.5281/zenodo.12179651) (2024).
- 306 **19.** Autorità Nazionale Anti Corruzione - Open Data website. ANACe. <https://dati.anticorruzione.it>.
- 307 **20.** Istituto Nazionale di Statistica. ISTAT. <http://www.istat.it/en>.
- 308 **21.** Fact Sheets on the European Union. Common classification of territorial units for statistics (NUTS). <https://www.europarl.europa.eu/factsheets/en/sheet/99/nomenclatura-comune-delle-unita-territoriali-statistiche-nuts->
- 309
- 310 **22.** Statistical codes of territorial administrative units: municipalities, metropolitan cities, provinces and regions. ISTAT.
311 <https://www.istat.it/it/archivio/6789>.
- 312 **23.** Main geographical statistics on municipalities. ISTAT. <https://www.istat.it/it/archivio/156224>.
- 313 **24.** Banca Dati Amministrazioni Pubbliche. Open BDAP. <https://openbdap.rgs.mef.gov.it>.
- 314 **25.** Italian Government. PNRR - National Recovery and Resilience Plan. <https://www.italiadomani.gov.it/content/sogei-ng/it/en/home.html>.
- 315
- 316 **26.** Gazzetta Ufficiale Repubblica Italiana. GURI. <https://www.gazzettaufficiale.it>.
- 317 **27.** Tenders Electronic Daily. TED. <https://ted.europa.eu/en>.
- 318 **28.** CPV codes and nomenclatures. <https://simap.ted.europa.eu/web/simap/cpv> (2022). Visited: 2022-12-01.
- 319 **29.** Batini, C. & Scannapieco, M. *Data and Information Quality: Dimensions, Principles and Techniques* (Springer, 2016).
- 320 **30.** Little, R. J. & Rubin, D. B. *Statistical Analysis with Missing Data* (John Wiley Sons, 2019).
- 321 **31.** Nai, R., Sulis, E., Pasteris, P., Giunta, M. & Meo, R. Exploitation and merge of information sources for public procurement
322 improvement. In *Machine Learning and Principles and Practice of Knowledge Discovery in Databases - International*
323 *Workshops of ECML PKDD 2022, Grenoble, France, September 19-23, 2022, Proceedings, Part I*, 89–102, [10.1007/978-3-031-23618-1_6](https://doi.org/10.1007/978-3-031-23618-1_6) (2022).
- 324
- 325 **32.** Nai, R. *et al.* AI applied to the analysis of the contracts of the italian public administrations. In *Proceedings of the Italia*
326 *Intelligenza Artificiale - Thematic Workshops co-located with the 3rd CINI National Lab AIIS Conference on Artificial*
327 *Intelligence (Ital IA 2023), Pisa, Italy, May 29-30, 2023*, 255–260 (2023).
- 328 **33.** Nai, R., Meo, R., Morina, G. & Pasteris, P. Public tenders, complaints, machine learning and recommender systems: a case
329 study in public administration. *Comput. Law Secur. Rev.* **51**, 105887, [10.1016/J.CLSR.2023.105887](https://doi.org/10.1016/J.CLSR.2023.105887) (2023).

330 Acknowledgements

331 The authors thank the Department of Management of the University of Turin for supporting this research with legal domain
332 experts in public tenders and, in particular, Prof. Gabriella Margherita Racca and Dr. Francesco Gorgerino for contributing data
333 analysis.

334 Author contributions statement

335 R.N. collected data, performed experiments, and wrote most part of the contribution; E.S. proposed the idea of building the DB,
336 and supervised the DB construction and article writing stages; R.M. supervised the initiative and maintained relations with the
337 institution.

338 **Competing interests**

339 The authors declare no competing interests.