Common European Data Spaces and Robust AI for Transparent Public Governance

CEDAR

Project acronym: CEDAR

Project full title: Common European Data Spaces and Robust AI for Transparent Public Governance

Call identifier: HORIZON-CL4-2023-DATA-01

Type of action: HORIZON-RIA

Start date: 01/01/2024

End date: 31/12/2026

Grant agreement no: 101135577

D2.2 Data Models and Preparation Tools V1

Document description: D2.2 provides the definition of the Data Model and indicators, as well as updates on the data collection and protection.

Work package: WP2

Author(s): Sophia Karagiorgou (UBI); Félix Cuadrado, David Rodríguez, José Miguel Blanco (UPM); Isabela Rosal Santos (KUL); Filippo Cassetti, Giacomo Delinavelli (ALBV); Mattia Tarchini (CNT); Stefanos Demertzis, Vasilis Poulos (CERTH); Jolanda Modic (ICS); Tjaša Kyovsky (MNZ)

Editor(s): Giulia Preti (CNT)

Leading partner: CNT

Participating partners: ALBV, CERTH, CNT, ICS, KUL, MNZ, UBI, UPM

Version: 1.2

Status: Final

Deliverable type: Report

Dissemination level: PU

Official submission date: 30/06/2025

Actual submission date: 30/06/2025

The CEDAR project has received funding from the European Union's Horizon Europe project call HORIZON-CL4-2023-DATA-01 funded project Grant Agreement no. 101135577

Disclaimer

This document contains material, which is the copyright of certain CEDAR contractors, and may not be reproduced or copied without permission. All CEDAR consortium partners have agreed to the full publication of this document if not declared "Confidential." The commercial use of any information contained in this document may require a license from the proprietor of that information. The reproduction of this document or of parts of it requires an agreement with the proprietor of that information.

The CEDAR consortium consists of the following partners:

No.	Partner Organisation Name	Partner Organisation Short Name	Country
1	Centre for Research and Technology Hellas	CERTH	Greece
2	Commissariat al Energie Atomique et aux Energies Alternatives	CEA	France
3	CENTAI Institute S.p.A.	CNT	Italy
4	Fundacion Centro de Technologias de Interaccion Visual y Comunicaciones VICOMTECH	VICOM	Spain
5	TREBE Language Technologies S.L.	TRE	Spain
6	Brandenburgisches Institut für Gesellschaft und Sicherheit GmbH	BIGS	Germany
7	Christian-Albrechts University Kiel	KIEL	Germany
8	INSIEL Informatica per il Sistema degli Enti Locali S.p.A.	INS	Italy
9	SNEP d.o.o	SNEP	Slovenia
10	YouControl LTD	YC	Ukraine
11	Artellence	ART	Ukraine
12	Institute for Corporative Security Studies, Ljubljana	ICS	Slovenia
13	Engineering – Ingegneria Informatica S.p.A.	ENG	Italy
14	Universidad Politécnica de Madrid	UPM	Spain
15	Ubitech LTD	UBI	Cyprus
16	Netcompany-Intrasoft S.A.	NCI	Luxembourg
17	Regione Autonoma Friuli Venezia Giulia	FVG	Italy
18	ANCEFVG – Associazione Nazonale Construttori Edili FVG	ANCE	Italy
19	Ministry of Interior of the Republic of Slovenia / Slovenian Police	MNZ	Slovenia
20	Ministry of Health / Office for Control, Quality, and Investments in Healthcare of the Republic of Slovenia	MZ	Slovenia
21	Ministry of Digital Transformation of the Republic of Slovenia	MDP	Slovenia
22	Celje General Hospital	SBC	Slovenia
23	Transparency International Deutschland e.V.	TI-D	Germany
24	Katholieke Universiteit Leuven	KUL	Belgium
25	Arthur's Legal B.V.	ALBV	Netherlands
26	DBC Diadikasia	DBC	Greece
27	The Lisbon Council for Economic Competitiveness and Social Renewal asbl	LC	Belgium
28	SK Security LLC	SKS	Ukraine
29	Fund for Research of War, Conflicts, Support of Society and Security Development – Safe Ukraine 2030	SU	Ukraine
30	ARPA Agenzia Regionale per la Protezione dell' Ambiente del Friuli Venezia Giulia	ARPA	Italy

Document Revision History

Version	Date	Modifications Introduced	
		Modification Reason	Modified by
0.2	09.05.2025	Added Section 3 and Subsection 2.2	Félix Cuadrado, David Rodríguez, José Miguel Blanco (UPM)
0.3	09.05.2025	Input under chapter Data Quality and Data Pseudonymization	Filippo Cassetti, Giacomo Delinavelli (ALBV)
0.4	15.05.2025	Added Section 2	Sofia Karagiorgou (UBI)
0.5	16.05.2025	Added Sections 4.3 and 4.4	Mattia Tarchini (CNT)
0.6	19.05.2025	Added Section 2.3	Stefanos Demertzis, Vasilis Poulos (CERTH)
1.0	22.05.2025	Added input in Section 2.2.3	Jolanda Modic (ICS)
1.1	30.05.2025	First pass after revision	Giulia Preti (CNT)
1.2	02.06.2025	Finalization	Giulia Preti (CNT)
2.0	30.06.2025	Final Version and submission	Thodoris Semertzidis, Mariana Minopoulou (CERTH)

Statement of Originality

This deliverable contains original unpublished work except where clearly indicated otherwise.

Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Table of Contents

1 Introduction 7
1.1 Positioning of the Deliverable within CEDAR 7
1.2 Structure of the Deliverable 7
2 Public and Private Data Collection and Synthetic Data Generation 9
2.1 New Dataset Collection 9
2.2 Data Quality and Compliance 9
2.2.1 Entity Deduplication 9
2.2.2 What is Data Quality from a Legal Perspective 10
2.2.3 Data for Slovenian Demo 12
2.3 Synthetic Data for the Slovenian Use Case 12
2.3.1 Architecture and Generation Process 13
3 Data Modelling, Harmonisation, Alignment 16
3.1 Rationale 16
3.2 Definition of the Data Model 17
3.2.1 Tender 17
3.2.2 Public Entity 19
3.2.3 Item 20
3.2.4 Bid 21
3.2.5 Organization 22
3.3 Simplifications for the MVP 22
3.4 Operations over the Data Model 23
3.4.1 Italian Pilot 23
3.4.2 Slovenian Pilot 23
3.4.3 Ukrainian Pilot 23
3.5 Formal Verification of the Data Model 24
4 Data Protection and (Pseudo)Anonymisation 25
4.1 Key Definitions: Personal Data, Anonymised Data, and Pseudonymised Data 25
4.2 Anonymisation vs Pseudonymisation 26
4.3 Implementation of Pseudonymisation in CEDAR 27
4.4 Description of the Pseudo-Anonymisation Tool for the MVP 28
4.4.1 Pseudo-Anonymisation Tool Pipeline 30
4.4.2 Description of the APIs 31
5 Conclusion 32
6 Bibliography 33

List of Figures

Figure 1: Data Quality dimensions. 11
Figure 2: Synthetic Generator Pipeline. 13
Figure 3: CEDAR Data Model. 17
Figure 4: Tender entity. 17
Figure 5: Tender relationship with Public Entity. 18
Figure 6: Tender relationship with Item. 19
Figure 7: Tender relationship with Bid. 19
Figure 8: Public_Entity entity. 20
Figure 9: Item entity. 20
Figure 10: Bid entity. 21
Figure 11: Bid relationship with Organization. 22
Figure 12: Organization entity. 22
Figure 13: Anonymisation vs Pseudonymisation. 26
Figure 14: A scheme of the Pseudo-anonymisation Tool Pipeline. 30

List of Tables

Table 1: Bids Template. 14
Table 2: Tenders Template. 15
Table 3: An example of Local Recoding. 29

List of Terms and Abbreviations	
API	Application Programming Interface
AUC	Area Under the ROC Curve
CJEU	Court of Justice of the European Union
DMP	Data Management Plan
DoA	Description of Actions
EDPS	European Data Protection Supervisor
ePO	eProcurement Ontology
GDPR	General Data Protection Regulation
I	Identifier attributes
LLM	Large Language Model
MAE	Mean Absolute Error
MSE	Mean Squared Error
MVP	Minimum Viable Product
OCDS	Open Contracting Data Standard
PRAM	Post RAndomization Method
RMSE	Root Mean Squared Error
S	Sensitive attributes
SRB	Single Resolution Board
UC	Use Case

QI	Quasi Identifier attributes

Executive Summary

The CEDAR project aims to provide high-quality, analytics-ready, and open data for transparent public governance. Deliverable 2.2 "Data Models and Preparation Tools V1" reports the second and third semesters of activities (M6-M18), including the consolidation of datasets to support pilot use cases, the definition and implementation of data quality best practices, the establishment of synthetic data generation for machine learning development and training, the design of a preliminary data model for knowledge graph development, and the implementation of data anonymisation techniques to ensure user privacy.

The report also outlines the project's approach to data quality within legal boundaries, details the data model's structure and the transparency indicators identified, presents validation of the data model, confirms GDPR compliance, and describes the tool developed to enforce these crucial data protection measures.

1 Introduction

The CEDAR project aims to address the critical need for high-quality, high-value, analytics-ready, and open data to enhance transparent public governance. Achieving truly transparent public governance, however, requires navigating the inherent tension between public access to information and the fundamental right to privacy. In fact, while transparency demands openness about government operations, releasing raw personal data can severely infringe on individual privacy. This deliverable, D2.2 Data Models and Preparation Tools V1, reports on key activities performed in the project's second and third semesters to bridge this gap.

First, we consolidated the inventory of data sources to establish the CEDAR Data Catalogue and identified the steps to ensure data quality from a legal perspective, assessed across six dimensions. In parallel, drawing on the literature review presented in Deliverable 2.1, we developed a tool for synthetic data generation to address data privacy restrictions, unavailability, or insufficiency.

Secondly, we defined a data model based on knowledge graph technologies to overcome the structural and semantic heterogeneity of collected datasets, enabling seamless data integration across domains, jurisdictions, and administrative systems. Within CEDAR, this knowledge graph encodes procurement processes, public bodies, economic operators, and contractual relationships as interconnected resources. To address the lack of transparency in public procurement, a key CEDAR goal, we identified potential indicators of lack of transparency and defined them as easily calculable operations over the data model. These indicators constitute clear purposes for the need to collect data from the data sources included in the inventory.

Finally, leveraging the knowledge from the privacy-preserving techniques literature review in Deliverable 2.1, we developed a REST API accessible tool to (pseudo)-anonymize structured and semi-structured textual data using non-perturbative methods like global and local recoding. This approach reduces disclosure risk while preserving the utility for statistical analysis and data mining.

These efforts directly contribute to enabling robust public access and analytical capabilities, while ensuring protection of user privacy and adherence to legal frameworks such as GDPR.

1.1 Positioning of the Deliverable within CEDAR

Building on the foundational work established in Deliverable 2.1, which detailed the initial data sources, preparation methodologies, and a comprehensive overview of synthetic data generation techniques explored during the project's first six months, this deliverable marks a significant step towards the practical implementation and application of these initial findings.

This deliverable reports on the consolidation of the identified datasets, the establishment of best practices for ensuring data quality from a legal perspective, the development and application of a synthetic data generation tool for the Slovenian pilot, the definition of a preliminary data model utilizing knowledge graph technologies, and the creation of a tool for data (pseudo-)anonymisation to address privacy concerns, all contributing to the overarching goal of providing high-quality, analytics-ready data for transparent public governance.

1.2 Structure of the Deliverable

This document continues with the following structure:

Section 2, "Public and Private Data Collection and Synthetic Data Generation", details the additional data sources identified by pilot users, the methodologies for ensuring data quality, and the process developed to generate synthetic data specifically for the Slovenian use case.

Section 3, "Data Modeling, Harmonisation, and Alignment", presents the data model and its formal verification, emphasizing the entities involved and their respective attributes.

Section 4, "Data Protection and (Pseudo)Anonymisation", addresses the ethical considerations and legal frameworks governing data handling, and describes the tool created to ensure adherence to these requirements.

Section 5, "Conclusion", provides a concise summary of the key findings and outcomes of this deliverable.

2 Public and Private Data Collection and Synthetic Data Generation

2.1 New Dataset Collection

To effectively monitor and analyse the impact of new public dataset collection of governmental contracts and decisions, we have collected the following data points in a relational database and an object storage aligned with CEDAR activities:

- **Unique Announcement Number:** The unique identifier assigned to each published governmental act or decision by the open data portal (in our case, the Greek one, DIAVGEIA).
- **Government Institution:** The specific governmental body (e.g., Ministry of Finance, Ministry of Health, Independent Authority for Public Revenue, etc.) responsible for issuing the act or decision.
- **Act/Decision Title:** A concise and descriptive title of the uploaded document.
- **Subject Matter/Keywords:** Relevant keywords or categories that describe the content and purpose of the act or decision.
- **Date of Issuance:** The official date when the act or decision was formally adopted by the issuing institution.
- **Date of Upload:** The exact date and time when the document was uploaded to the "Transparency Portal."
- **Document Format and Digital Content:** The file format and the actual file (e.g., JSON, PDF, DOCX, TXT) in which the act or decision is published.
- **Digital Signature Status:** Confirmation of the presence and validity of the digital signature associated with the document.

The process of data collection has been performed through the development of scheduled tasks designed to automatically extract information from the DIAVGEIA Portal. The acquired data are systematically stored within a relational database as well as an object storage system. The indexing of this information facilitates the ability to search through the documents and precisely discern specific governmental information and documents based on multiple criteria (e.g., time, type of ministry, keywords, etc.), irrespective of the technical expertise of the end user.

This dataset allows for the evaluation of compliance by government institutions with national and EU legislation[1] and the online publication mandate. It also permits the extension of research activities beyond the CEDAR pilot tasks and enables the examination of other public datasets, aligned with detecting patterns or anomalies that may suggest delays in publication, inconsistencies in data management, or potential issues of concern related to governmental activities, national security, or other matters. Finally, by analysing the relationship between online publication and the indicators of transparency, corruption, and citizen engagement, it becomes possible to identify areas for improvement in the publication process, contributing to enhanced governmental transparency.

2.2 Data Quality and Compliance

Ensuring high data quality is a fundamental prerequisite for any system that aims to analyse, correlate, or extract insights from heterogeneous data sources.

2.2.1 Entity Deduplication

The CEDAR project is collecting a data catalogue, as described in D2.1, with a curated collection of data sets from each of the participating pilots. These data sources are highly heterogeneous, differing in the source organisation, language, and scope. These differences in national data practices often result in inconsistencies, missing values, and naming variations. One frequent challenge is entity matching, where the same real-world entity (such as a company) may appear under slightly different names or incomplete identifiers across datasets. If not appropriately addressed, such inconsistencies compromise the reliability of higher-level analyses, such as financial tracking, fraud detection, and transparency assessments.

Within the Slovenian pilot, we addressed an **entity deduplication task** involving three datasets:

- *JN SBC*: Companies awarded public tenders, without registry numbers.

[1] http://elib.aade.gr/elib/view?d=/gr/act/2013/4210

- *All Companies in SLO*: A national registry containing company names, registry numbers (matična številka), and addresses.
- *SBC TRX ALL*: Records of financial transactions involving public bodies.

The objective was to link companies from *JN SBC* to their corresponding entries in *All Companies in SLO* or, if necessary, in *SBC TRX ALL*, and to retrieve the associated registry numbers. This information was needed to enable accurate aggregation of financial data, monitoring of contract awarding practices, and detection of irregularities.

The cleaning process involved several steps:

1. Preprocessing: Company names were normalized by lowercasing, removing diacritics, eliminating legal suffixes (such as "d.o.o." or "d.d.") via regular expressions, removing punctuation, and discarding common company-related tokens (e.g., "Ltd", "Inc."). This preprocessing minimized artificial differences between records.
2. String Matching: Names were compared using a combination of Levenshtein distance and token set ratio metrics, assigning a higher weight to the similarity of the first and second words, which typically carry the most distinctive information in Slovenian company names.
3. Acceptance Criteria: Matches were accepted if the similarity score reached or exceeded 90%. Otherwise, the best available match was proposed for manual review.
4. Validation: Ambiguous results were flagged for manual inspection.

The methodology achieved a match rate above **92%**, significantly improving the consistency and interpretability of the pilot datasets. While the outcome of this task has been applied to the three mentioned data sources from the Slovenian pilot, the techniques applied are applicable to other cases where entity reconciliation across heterogeneous data sources is required.

2.2.2 What is Data Quality from a Legal Perspective

This section explores the concept of data quality within the legal frameworks that govern data processing, particularly under the General Data Protection Regulation (GDPR) and other pertinent European regulations.

Maintaining a high standard of data quality is an important consideration within the CEDAR project, as it enhances the consistency and interpretability of results and supports the broader aims of the research. From a legal perspective, data quality does not have an explicit legal definition. The European Data Protection Supervisor underscores that data quality should be interpreted within the broader framework established by Article 5 of the General Data Protection Regulation (GDPR). This provision outlines fundamental data processing principles, including lawfulness, fairness, accuracy, and transparency. It also introduces the purpose limitation principle, requiring that personal data be collected only for specified, explicit, and legitimate purposes. Closely related is the principle of data minimisation, which stipulates that only data necessary for those purposes should be processed. Lastly, the principle of accuracy obliges data controllers to ensure that personal data is accurate and up to date and that any inaccuracies are rectified without delay. Collectively, these principles safeguard individual rights and increase the integrity and trustworthiness of data processing activities.

Data quality is also mentioned in the Handbook on European Data Protection Law, issued by the Council of Europe. It emphasises that data must be adequate, relevant, and not excessive (following the principle of proportionality), and it also highlights the necessity for data to be accurate. According to the Handbook, the data quality principle necessitates that controllers take measures to correct any inaccuracies, mitigate the risks associated with profiling, and periodically assess the quality of the data and the algorithms employed.

In 2021, the Publications Office of the European Union issued the Data Quality Guidelines, providing recommendations and best practices for producing and sharing high-quality data. The document emphasises the importance of providing such data in a context that experiences an increasing demand for high-quality publicly available data that can be reused for different purposes. Low-quality data severely limits the potential for reuse, as some datasets may be incomprehensible due to poorly defined or inaccurate components, such as missing values, inconsistencies, undefined data types, and inadequate documentation regarding their structure or formats (such as HTML, GIF, or PDF). The Guidelines suggest that data should be easily discoverable, analyzable, and visualizable to ensure that data is readily reusable. Reusers should clearly understand what data represents, their definitions and structure, ideally provided in their preferred format. According to the Guidelines, it's recommended that data be processed based on the FAIR principles—findability, accessibility, interoperability, and reusability—which provide a helpful framework for organising data quality elements. This is also sustained in Art. 10 Open Data Directive for 'research data'. The CEDAR Project embraces those suggestions and is elaborated in the Data Management Plan (D7.2 issued on 28.06.2024). Developing a data management plan addresses critical questions regarding how to collect and clean data, where to publish data, how to store metadata, which formats to use, and which standards to adhere to, thereby facilitating the publication process.

However, in the absence of an explicit legal definition, academics have examined data quality and emphasised that it involves much more than simply accuracy. They argue that high-quality data is defined as being "fit for purpose," which extends well beyond just the accuracy of the information. This is because different users have varying requirements and assess data quality based on their specific needs and the intended purposes of processing the information. Accuracy is one characteristic, but it does not guarantee data quality. The overall consideration of other characteristics, such as format and structure, can make the data less fit for specific processing purposes. Data considered accurate should represent real-world facts and be suitable for informed decision-making. Therefore, to evaluate the quality of data, it is common to refer to multiple characteristics of data, also known as dimensions, which indicate the overall quality level. Despite extensive literature and ongoing discussions about data quality dimensions, a consensus has yet to be reached on which of them are essential for assessing data quality since those depend on the purposes and the context of the processing. The relevance and significance of these dimensions can vary, differing across organisations and data types.

The most used dimensions to assess data quality are:

- **Completeness**: The extent to which data is sufficiently broad, deep, and comprehensive for its intended use.
- **Accuracy**: The degree to which data is correct, reliable, and verified.
- **Timeliness**: This dimension refers to the readiness of the data within a given time frame.
- **Consistency**: The degree to which data is presented uniformly and is compatible with previous datasets.
- **Uniqueness**: Ensures the dataset is free from duplicate records and that every record can be uniquely identified.
- **Validity**: The extent to which data adheres to defined business rules and parameters, ensuring proper structure and expected values.

Accuracy
Consistency
Validity
Data Quality
Completeness
Timeliness
Uniqueness

Figure 1: Data Quality dimensions.

Furthermore, we need to emphasise that data quality is not a fixed or static concept but a dynamic and evolving one. Quality changes throughout the data lifecycle, from collection to processing, to eventual use. Various stages of handling data can introduce specific issues, such as measurement inaccuracies during data gathering, mistakes during data entry, or subjective interpretations when presenting the information. This means that the data quality at the collection point may differ significantly from its quality when structured, displayed, or used in decision-making. These considerations are important not only for technical reasons but also for legal ones. As mentioned above, in the European Union, data protection laws already include rules related to data quality. This means that people should have the right to ensure that companies handle their personal data in a way that meets these quality standards, such as ensuring the data is accurate and up to date (accuracy), only collected for specific and legitimate purposes (purpose limitation), limited to what is necessary (data minimisation), and not kept longer than necessary (storage limitation), as established in the above-mentioned Article 5 of the General Data Protection Regulation (GDPR).

In CEDAR's area of work, having a clear data governance policy can help ensure data quality throughout the entire process of managing and handling data. Data governance is a set of principles, standards, and practices. They apply to the end-to-end data lifecycle (collection, storage, use, protection, archiving, and deletion). Data governance policy needs an implementation, and this is carried out through the Data Management Plan (D7.2 published on 01.07.2024), which is the set of rules and guidelines that explain how data should be collected, stored, used, and shared. For example, it can outline the methods to gather data to ensure it is accurate and complete. It also helps determine what data should be kept, for how long, and who is responsible for ensuring everything is done properly. These policies also set out what standards and best practices should be followed when processing data, which in turn helps to improve the overall quality of the data.

2.2.3 Data for Slovenian Demo

To ensure that the data used in the Slovenian Demo is accurate and timely, we applied a mix of technical and procedural safeguards. These include deduplication, normalisation, format standardisation, and semantic harmonisation, ensuring consistency across diverse sources. Specifically, duplicated entries were removed, formatting inconsistencies were corrected, and invalid or incomplete records were identified and excluded. Logical checks were applied to ensure consistency across fields. Actions are in line with the applicable law, guaranteeing lawful, transparent, and fair data processing. Data handling respects principles of accuracy, minimisation, and purpose limitation under the GDPR Article 5.

2.3 Synthetic Data for the Slovenian Use Case

CERTH developed a generator pipeline to create a synthetic dataset to support technical validation and address data availability challenges without relying on proprietary or sensitive real-world procurement information. The generator produces both regular and potentially fraudulent examples that mirror the structure, semantics, and variability of actual procurement data, based on insights extracted from the Slovenian pilot.

The synthetic data generator was developed in Python, using the Faker library and a large language model (LLM). Faker generates simple synthetic values such as names, addresses, professional email addresses, domain names, and tax numbers, providing structurally realistic data without exposing personally identifiable information or proprietary content. For more complex or longer fields, data is generated using an LLM, specifically Gemma 2 27b, via prompt engineering techniques on an Ollama server running two NVIDIA RTX 4070 GPUs with 12GB of VRAM each. These complex fields include tender technical specifications, general requirements, product or service names, and numerical values dependent on multiple parameters. Examples include the number of certified employees, quantities of goods and services, median market prices, technical specifications, and unique trademarks. A detailed breakdown of simple and complex fields is provided in Tables 1 and 2.

To illustrate what is meant by complex fields with multiple parameters, we provide two prompt examples: the **median price of goods** and **service requirement descriptions**. These examples demonstrate how generating such values can depend on the internal knowledge and reasoning capabilities of the LLM. Rather than deriving these values through a manual analysis of external data, we use prompt engineering to guide the LLM in making educated approximations. These approximations are considered complex because they depend on various contextual factors such as the type of good, its technical specifications, industry standards, or service complexity, which would otherwise require extensive domain-specific datasets to model accurately.

For **unit price estimation**, the prompt is:

“In a public procurement, the product: '{product_name}', offered as a bid in a tender. What is a reasonable price per unit in EUR when offering {quantity} units? If you cannot provide specific pricing information, just make a reasonable guess. Provide only the price per unit in EUR, without any additional text.”

The model returns a single numeric value, which is internally post-processed to simulate price variation across suppliers and markets.

For **service requirement descriptions**, the prompt is:

“Write a short single-line paragraph as tender issuer {issuer_name}. You are a medical institution that needs to purchase a service of {service_name}. The general requirements are {requirements}. You need {number} certified employees to be provided by the bidding company. Refer to 'we' not 'I'. Do not refer to your company. The paragraph must be text as it is written to the tender. Do not provide lists, new lines, or placeholders. Add some creativity.”

This produces varied, natural-sounding requirement texts.

2.3.1 Architecture and Generation Process

The pipeline comprises two parallel branches for generating **regular** and **potentially fraudulent** procurement scenarios. Each branch contains four core modules, executed sequentially. A visual overview of these components follows, providing a representation of the modular structure and data flow within the generation process.

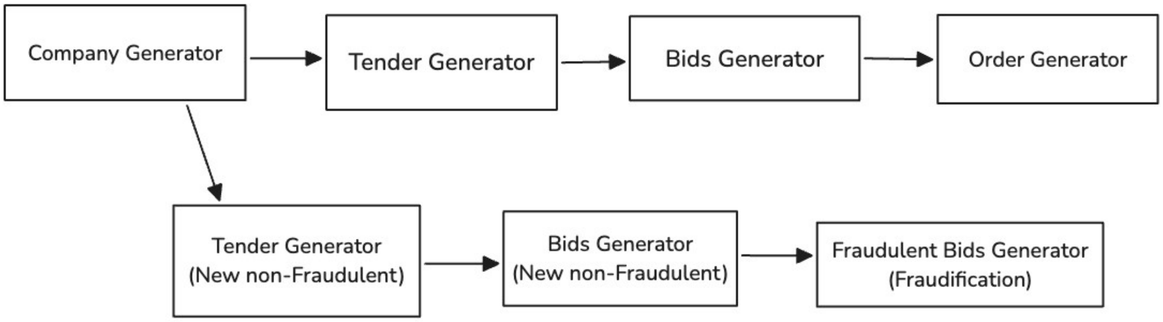

Company Generator
Tender Generator
Bids Generator
Order Generator
Tender Generator (New non-Fraudulent)
Bids Generator (New non-Fraudulent)
Fraudulent Bids Generator (Fraudification)

Figure 2: Synthetic Generator Pipeline.

2.3.1.1 Regular Data

A total of 600 tenders were generated to represent typical, non-fraudulent scenarios along with their corresponding bids, orders, and company information. These examples follow standard procurement patterns and do not include any risk signals or anomalous behaviour.

2.3.1.2 Potentially Fraudulent Data Injection

To enable the development and evaluation of risk detection methodologies, the second part of the pipeline simulates potentially fraudulent procurement scenarios by injecting known anomalies and irregularities based on four key Slovenian Pilot fraud indicators: number of bids (both absolute and relative), price, date, and requirements. These indicators reflect potential misconduct such as coordination or bias.

Specifically, the fraud cases implemented are:

- **Single-Bid Tenders**: Only one submission per tender (absolute bid number anomaly).
- **Bid Volume Irregularities**: Unusually high numbers of bids (10 to 15) compared to typical historical data (3 to 6) (relative bid number anomaly).
- **Suspicious Price Offers**: Winning bids with disproportionately low prices compared to competitors.
- **Date Anomalies**: Bids submitted unusually close to the publication or deadline dates, suggesting possible coordination.

Each type of fraud is represented by 20 distinct instances, resulting in a total of 100 potentially fraudulent tenders. These are clearly labelled and exported separately from the regular dataset to enable focused analysis and validation.

2.3.1.3 Output Data and Formats

The dataset consists of 15 tabular .csv files. For the regular dataset, 2 files are the Slovenian translation of the generated tenders and bids.

- synthetic_tenders_3.csv
- synthetic_bids_3.csv
- synthetic_orders.csv
- ...
- synthetic_tenders_3_sl.csv
- synthetic_bids_3_sl.csv

For the potentially fraudulent data, 10 files are for tenders and bids, 4 of each labelled with their anomaly indicator, while files marked with "_total" contain the 100 potentially fraudulent tenders and their corresponding bids.

- synthetic_bids_fraudulent_number_of_bids_absolute.csv
- synthetic_bids_fraudulent_number_of_bids_relative.csv
- synthetic_bids_fraudulent_prices.csv
- synthetic_bids_fraudulent_timing.csv
- synthetic_bids_fraudulent_total.csv
- ...
- synthetic_tenders_fraudulent_number_of_bids_absolute.csv
- synthetic_tenders_fraudulent_number_of_bids_relative.csv
- synthetic_tenders_fraudulent_prices.csv
- synthetic_tenders_fraudulent_timing.csv
- synthetic_tenders_fraudulent_total.csv

Table 1: Bids Template.

Field Name	Gen Type	Description
bid_id	Non-LLM	Unique identifier for the bid.
tender_id	Non-LLM	Identifier of the tender this bid is related to.
company_name	Non-LLM	Name of the bidding company.
company_address	Non-LLM	Address of the bidding company.
bidder_tax_number	Non-LLM	Tax identification number of the bidding company.
date_of_issue	Non-LLM	Date the bid was issued.
prepared_by	Non-LLM	Name of the person who prepared the bid.
offered_items	LLM	List of item names offered in the bid.
offered_items_descriptions	LLM	Descriptions of the items offered.
offered_items_technical_specs	LLM	Technical specifications of each offered item (goods specific value).
offered_quantities	Non-LLM	Quantities for each item offered.
offered_prices	LLM	Unit prices for each item offered.
offered_discounts	None	Discounts applied per item, if any.
total_price	Non-LLM	Total price before discount.
total_discount	None	Total discount amount.
final_price	Non-LLM	Final total price after applying discounts.
is_valid	LLM-judged	Whether the bid is valid (The offer meets the tender's requirements)

is_winner	Non-LLM	Whether this bid was selected as the winning bid.

Table 2: Tenders Template.

Field Name	Gen Type	Description
tender_id	Non-LLM	Unique identifier for the tender.
tender_type	Non-LLM	Type of tender. (Service or Goods)
tender_issuer	Non-LLM	Organization issuing the tender.
service_good_name	LLM	Name of the service or good requested.
quantity	LLM	Required quantity of the service/good.
num_certified_employees	LLM	Required number of certified personnel.
general_requirements	LLM	List of general requirements.
general_requirements_text	LLM	General requirements in plain text.
request_for_delivery	LLM	Whether delivery is requested.
request_for_installation	LLM	Whether installation is requested.
request_for_user_training	LLM	Whether user training is required.
request_for_maintenance	LLM	Whether maintenance is required.
request_for_certificates	LLM	Whether official certifications are required.
technical_requirements	LLM	List of technical requirements.
contact_persons	Non-LLM	List of contact persons for the tender.
publication_date	Non-LLM	Date the tender was published.
deadline	Non-LLM	Submission deadline for the tender.
prepared_by.name	Non-LLM	Name of the person who prepared the tender document.
prepared_by.email	Non-LLM	Email of the person who prepared the tender document.

3 Data Modelling, Harmonisation, Alignment

3.1 Rationale

The CEDAR data model is based on knowledge graph technologies to address the structural and semantic heterogeneity of the collected datasets. Knowledge graphs provide a flexible and expressive modelling paradigm that supports the representation of entities, relationships, and contextual metadata in a unified framework. This is particularly suited to scenarios like those addressed in CEDAR, where data from multiple domains, jurisdictions, and administrative systems must be integrated.

The adoption of a graph-based data model aligns with architectural principles promoted by the European Strategy for Data and the technical foundations of the Common European Data Spaces. Specifically, knowledge graphs offer native support for Linked Data principles and semantic reasoning, which are instrumental for enabling interoperability across data spaces and facilitating data federation at the European level. We will further work on the project in the alignment of data imported and exported through Data Spaces so that it can be related to the concepts presented in this model.

Within CEDAR, the knowledge graph enables the encoding of procurement processes, public bodies, economic operators, and contractual relationships as interconnected resources. This facilitates entity resolution, enrichment from external sources, and the definition of high-level indicators relevant to the domain of public procurement integrity. Furthermore, the model can be incrementally extended to cover additional use cases and data domains without disrupting existing structures. We have worked on an initial version of the data model that supports every requirement defined by the three CEDAR pilots in the definition of the MVP, and the initial use cases supported by the architecture. This serves as a validation of the methodology and a base point for further iteration of this model into the full CEDAR specification.

As the approach is bottom-up, starting from the use cases, we are not strictly adhering to a specific vocabulary or ontology, such as the eProcurement ontology (ePO), or the Open Contracting data model. The main reason behind this is twofold: On the one hand, these ontologies and data models are static structures that lack the flexibility that is required for the CEDAR data model; they would require constant and detailed updates to capture the specifics of CEDAR. On the other hand, while they are extensive and are able to capture most common situations, they would require extensive rework to capture certain topics that are discussed in detail within CEDAR. Furthermore, it would also be necessary to take out multiple elements of these structures that are not and will not be considered in CEDAR. Nonetheless, our abstractions can be aligned to standardisation initiatives like eProcurement Ontology (ePO), Digiwhisk, or the Open Contracting Data Standard (OCDS). This alignment can happen because the CEDAR data model constitutes a subset of certain instances of these initiatives, but also because certain aspects are checked to ensure compliance. Moreover, CEDAR data model instances can be serialised using the JSON-LD standard, ensuring interoperability with existing and emerging components of the European data infrastructure.

3.2 Definition of the Data Model

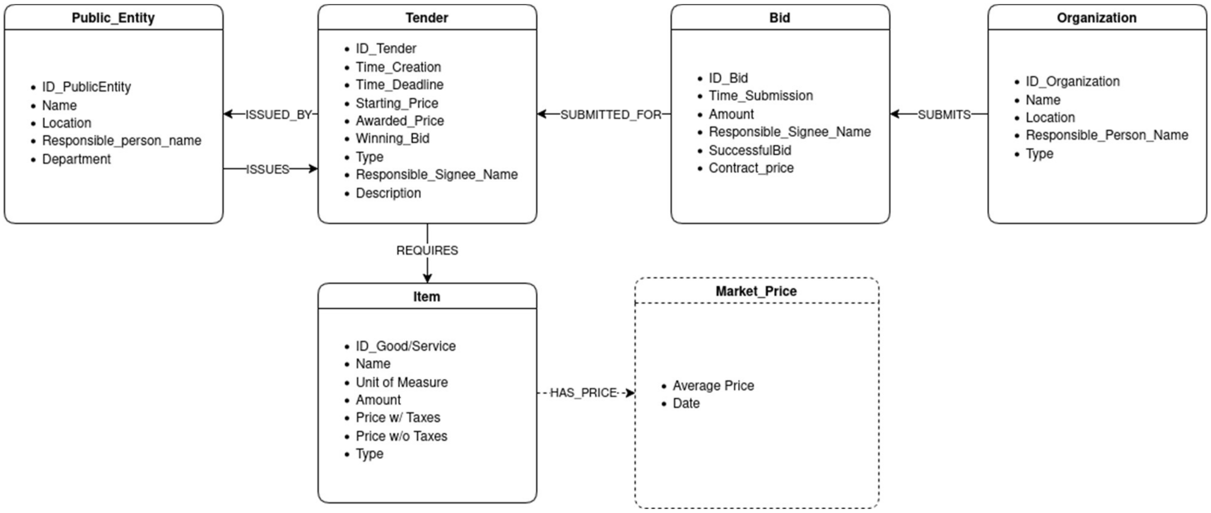


Figure 3: CEDAR Data Model.

Figure 2 displays the data model defined for the CEDAR MVP. This data model is comprised of 5 entities, 32 attributes, and 5 relations. To understand it, we will describe each of the above entities, attributes, and relations in detail. For that matter, we will go over all the entities, their attributes, and the relations between entities.

Let it be stated that it is possible that, in the context of the MVP, fields of the data model remain unpopulated in the CEDAR Knowledge Graph. This is because the data model is created to encompass the multitude of use cases that have been selected towards the MVP, and that leads to some fields not being present in all of them.

3.2.1 Tender

The entity **Tender** is the central entity of the initial knowledge graph model. A tender represents a formal request to supply goods or provide a task, service, or job. For a tender, different organisations will submit bids, including an offered price.

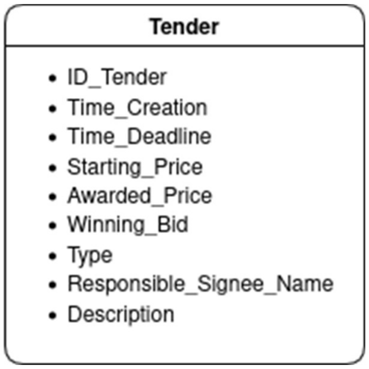


Figure 4: Tender entity.

This entity is characterized by the attributes *ID_Tender*, *Time_Creation*, *Time_Deadline*, *Winning_Bid*, *Type*, *Responsible_Signee_Name,* and *Description*, among others.

- *ID_Tender* is a unique identification code for the tender, assigned by the organisation issuing it. This code is an undetermined number of alphanumeric characters.
- *Time_Creation* is the date and time of the creation of the tender by the public entity. This attribute will have a DateTime type.

- *Time_Deadline* is the date and time of the end of the tender, the deadline for submitting any bids to the tender. This attribute will have a DateTime type.
- *Starting_Price* is the initial price set when creating the Tender to provide a reference for the submitted bids.
- *Awarded_Price* is the price that will be paid by the winning bid.
- *Winning_Bid* is the unique identification code for the bid submitted by an organisation that will supply the services or goods required for the tender. This code is an undetermined number of alphanumeric characters.
- *Type* is the characterization of the tender with respect to what is required to be supplied. In the given context of the MVP, it can be Goods, Services, or Both.
- *Responsible_Signee_Name* is the full name of the person within the public entity who is responsible for the publication, fairness, and well-being of the tender. It is an undetermined number of alphabetical characters.
- *Description* is a detailed listing and explanation of the goods and services required by the tender, as well as of the context in which the tender happens. It is an undetermined number of alphanumeric characters.

Tenders have three types of relationships with other types of entities from the model, to capture the underlying authorities, goods or services that must be supplied, and the received bids. We briefly describe each relation in the following lines.

First, every Tender has a one-to-one relationship with a Public Entity. The Public Entity provides details about the organisation that issues the Tender.

The relationship is characterised by the *ID_Tender* and *ID_PublicEntity* properties of each Entity.

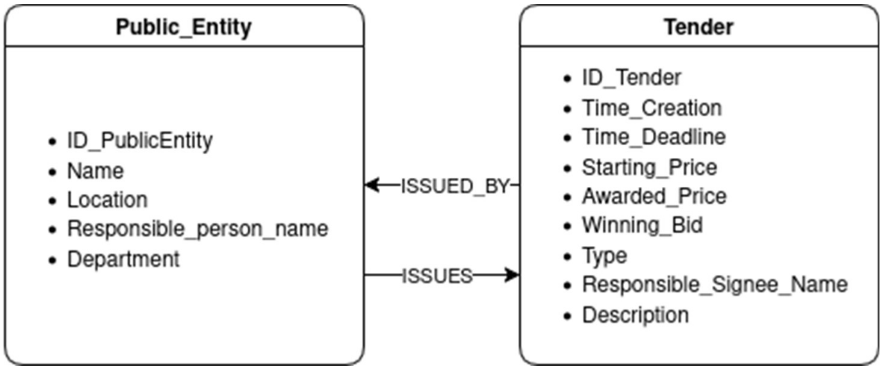


Figure 5: Tender relationship with Public Entity.

Second, Tenders have a relation with one Item entity. This relation is based on the idea that any tender is going to require several items, whether goods or services. In the initial version of the MVP, it is expected that a tender will only require one type of item (with multiple relationships being possible for more complex Tender definitions).

The relationship is characterised by the *ID_Tender* and *ID_Good/Service* properties of each Entity.

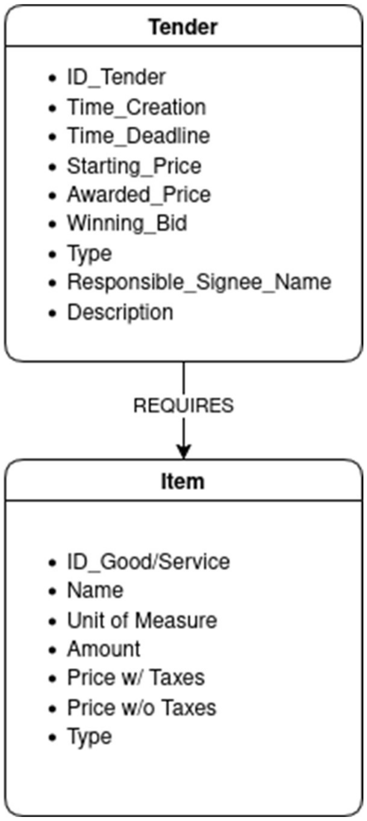


Figure 6: Tender relationship with Item.

Third, Tenders have a relation with one or more Bids. The relation links each Bid to the corresponding Tender. In particular, any tender is expected to receive several bids that do their best to fulfil the needs of the tender.

The relationship is characterised by the *ID_Tender* and *ID_Bid* properties of each Entity.

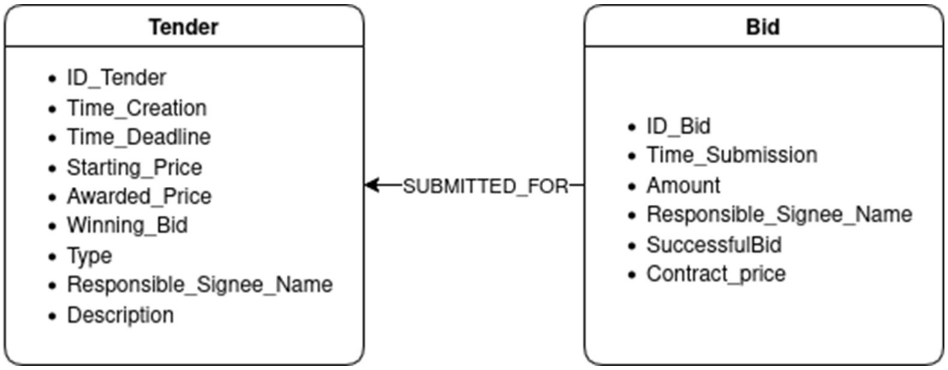


Figure 7: Tender relationship with Bid.

3.2.2 Public Entity

The entity **Public_Entity** is the representation of an organisation or body providing services to the public on behalf of the government. In the case of CEDAR, it will be the organisation that is issuing the tender, on which later the participants will bid.

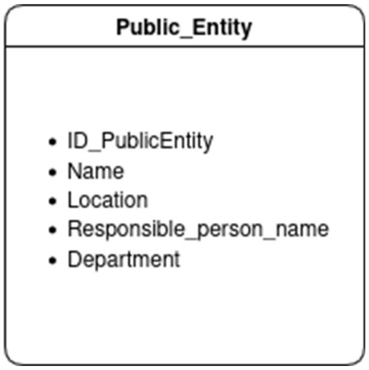


Figure 8: Public_Entity entity.

This entity is characterised by the attributes *ID_PublicEntity*, *Name*, *Location*, *Responsible_person_name*, and *Department*.

∉ *ID_PublicEntity* is a unique identification code for the organisation. This code is an undetermined number of alphanumeric characters.

∉ *Name* is the name of the public entity that issues the tender. This name might not be unique. It is an undetermined number of alphanumeric characters.

∉ *Location* is the address where the headquarters of the public entity are located. It is an undetermined number of alphanumeric characters.

∉ *Responsible_person_name* is the full name of the person within the public entity that authorises tenders. It is an undetermined number of alphabetical characters.

∉ *Department* is the suborganisation within the public entity that is issuing the public tender. It is an undetermined number of alphanumeric characters.

3.2.3 Item

The entity **Item** is the representation of a request for one type of goods or services within a Tender. This entity is characterised by having the attributes *ID_Good/Service*, *Name*, *Unit of Measure*, *Amount*, *Price w/ Taxes*, *Price w/o Taxes*, and *Type*.

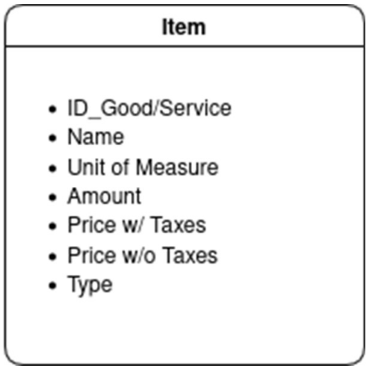


Figure 9: Item entity.

This entity is characterised by the following attributes:

∉ *ID_Good/Service* is the unique identification code of the good/service that is expected to be supplied for the tender. This code is an undetermined number of alphanumeric characters.

∉ *Name* is the name of the item that is required for the bid. This name might not be unique. It is an undetermined number of alphanumeric characters.

- ∉ *Unit of Measure* is the definite magnitude used to measure the item that is required for the tender. It is an undetermined number of alphanumeric characters.
- ∉ *Amount* is the number of items that are required by the tender. It is a number.
- ∉ *Price w/ Taxes* is the cost of one item, including all the corresponding taxes and expressed in euros. It is a number.
- ∉ *Price w/o Taxes* is the cost of one item without including any of the corresponding taxes and expressed in euros. It is a number.
- ∉ *Type* is the specification of the kind of item that will be required for the tender: either an item or a good. It is an undetermined number of alphabetic characters.

3.2.4 Bid

The entity **Bid** is the representation of the offer that one organisation makes to fulfil the requirements of the tender issued by the public entity. In particular, this bid will offer a certain amount of the services/goods needed by the original tender at a certain price.

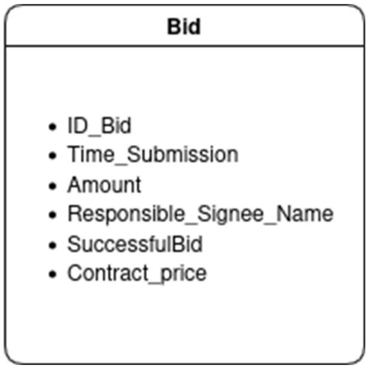


Figure 10: Bid entity.

This entity is characterised by having the attributes *ID_Bid*, *Time_Submission*, *Amount*, *Responsible_Signee_Name*, *SuccessfulBid*, and *Contract_price*.

- ∉ *ID_Bid* is the unique identification code for the bid, submitted by an organisation which will supply the services or goods required for the tender. This code is an undetermined number of alphanumeric characters.
- ∉ *Time_Submission* is the date and time of the submission of the bid by the organisation. This attribute will have a DateTime type.
- ∉ *Amount* is the number of goods/hours of service that the bid is contributing to those expected by the tender. This may differ from the required by the tender. It is a number.
- ∉ *Responsible_Signee_Name* is the full name of the person within the organisation who submits the bid, who is responsible for the publication, fairness, and well-being of it. It is an undetermined number of alphabetical characters.
- ∉ *SuccessfulBid* is the indicator of the success of the bid in winning the tender. It will be a boolean stating if the organisation won (or not) the tender.
- ∉ *Contract_price* is the amount that will be charged to the public entity for the provision of the goods/services required, expressed in euros. It is a number.

There is a one-to-one relationship between Bid and Organization. This allows us to establish a relation between the known data about the organisation and the specific bid submitted for the tender that the Public Entity has issued. It is expected that an organisation will only issue one bid for the same tender, but multiple bids aimed at different tenders.

The relationship is characterised by the *ID_Bid* and *ID_Organization* properties of each Entity.

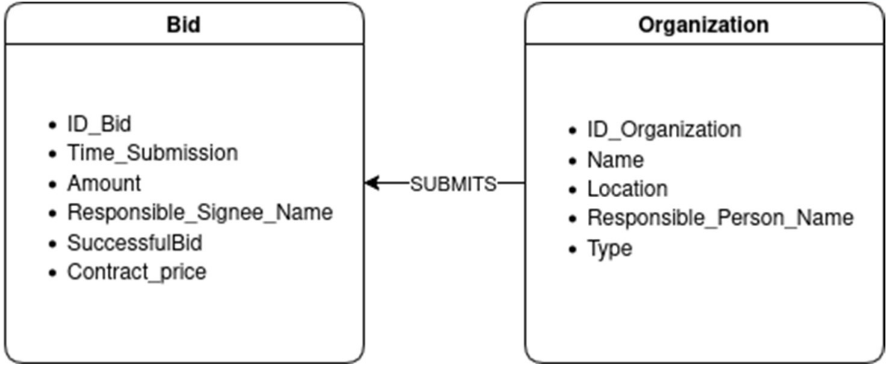


Figure 11: Bid relationship with Organization.

3.2.5 Organization

The entity **Organization** is the representation of the entity that submits the bid for the tender issued by a public entity. This organisation is aimed to be a private company or a conglomerate of them.

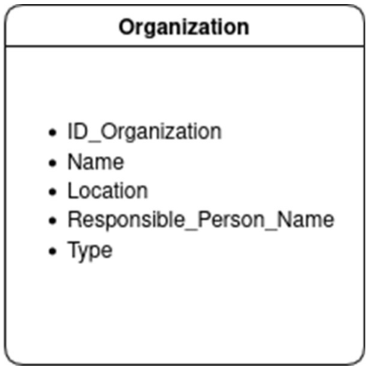


Figure 12: Organization entity.

This entity is characterised by having the attributes *ID_Organization*, *Name*, *Location*, *Responsible_Person_Name*, and *Type*.

∉ *ID_Organization* is a unique identification code for the organisation. This code is an undetermined number of alphanumeric characters.

∉ *Name* is the name of the organisation that submits the bid. This name might not be unique. It is an undetermined number of alphanumeric characters.

∉ *Location* is the address where the headquarters of the organisation are located. It is an undetermined number of alphanumeric characters.

∉ *Responsible_Person_Name* is the full name of the person within the organisation who authorises bids. It is an undetermined number of alphabetical characters.

∉ *Type* is the kind of organisation that submits the bit. It could be a private company, a conglomerate, or any other type of organisation that might be able to submit a bid. It is an undetermined number of alphabetical characters.

3.3 Simplifications for the MVP

The current data model of CEDAR is not the final version, but rather a smaller version targeted for the MVP. With that in mind, it is important to understand that the current state of the data model has been designed to cater to the use cases that were selected in cooperation with the pilots and only those. This means that there are parts of the data collected by the pilots that are currently receiving no coverage in the data model, and while that would be an issue, it is not in the current state of the project.

The data model will be extended to ensure that it provides full coverage of the data provided by the pilots, as well as any possible extensions that will be hinted at during the explanation and definition of further use cases. The full version of the data model is expected to be created from the full data available and constant feedback from both the pilots as well as other partners involved in similar topics, such as the knowledge graph that the data model supports.

3.4 Operations over the Data Model

The Italian, Slovenian, and Ukrainian pilots identified several indicators of a lack of transparency. These indicators are defined over the data model as operations in such a way that they can be calculated easily. These indicators are part of those defined in D5.1. Demonstration Preparation Report v1.0, and there, the full explanation of each of them can be found.

3.4.1 Italian Pilot

In the Italian Pilot UC#2, negotiated procedures are specifically scrutinised. These procedures can also be started for three different kinds of assets: works, goods, and services. Therefore, n*egotiated procedures* will be modelled as **Tender** entities with the attribute *Type* specified as "negotiated", joint with one of the three kinds ("works", "goods", "services").

UC#2-AI#1: NUMBER OF Negotiated procedures for a single internal unit

This indicator is calculated by measuring the number of ISSUES relations where the same **Public_Entity** entity (as determined by the id) is connected to **Tender** entity with the attribute *Type* specified as "negotiated". Different internal units will have unique **ID_Tenders**.

UC#2-AI#2: NUMBER OF Negotiated procedures for item category

This indicator is calculated by measuring the number of ISSUES relations of a **Public_Entity** entity connected to **Tender** entity with the attribute *Type* specified as one of the three kinds of "negotiated".

UC#2-AI#3: List of invited companies and received bids

This indicator is determined by calculating the ratio of the number of SUBMITS relations of **Organization** with *Type* set to "invited" that connects to a **Bid** that is SUBMITTED_FOR for a **Tender.** To compare with non-invited bids, it is possible to register the number of SUBMITS relations of **Organization** with *Type* set to any other value. These organisations will connect to a **Bid** that is SUBMITTED_FOR for the same **Tender** with the same *ID_Tender*. Note that the data model does not explicitly capture the concept of an invited organisation to submit a bid, so this would be a possible workaround by making use of the *Type*. Another alternative would be to include the full list of invited organisations as part of the Description (or even their number), which is already free text, and use that to compute the ratio rather than modelling relationships that express invitation and not submission.

3.4.2 Slovenian Pilot

UC#5-AI#1: NUMBER OF BIDS: (Relative) number of submitted bids for a specific good/service in historical data

This indicator is calculated by first selecting a specific **Item** (characterised by the *ID_GoodService*). Then, for each **Tender** that has a REQUIRES relationship with the entity, we measure the number of SUBMITTED_FOR relations connecting **Bid** entities to the **Tender**. From there, statistical analysis per unique item can be performed.

UC#5-AI#2: NUMBER OF BIDS: (Absolute) number of submitted bids for a specific tender

This indicator is calculated by measuring the number of SUBMITTED_FOR relations connecting **Bid** to **Tender** with the same *ID_Tender*.

UC#5-AI#3: REQUIREMENTS: Too strict/limiting/specific requirements can be defined to exclude other bidders

This indicator is calculated by researching the text included in *Description* of all the **Tender** with the same *Type*.

3.4.3 Ukrainian Pilot

UC#2-AI#5: The ultimate beneficial owner (controller) of the legal entity is its chief executive or signatory

This indicator is calculated by extracting the *Responsible_Person_Name* of the **Organization**.

UC#5-AI#1: Rebuild Ukraine procurement: winning bids

This indicator is calculated by extracting the *Winning_Bid* of Tender related to the SUBMITTTED_FOR **Bid**.

UC#5-AI#4: Rebuild Ukraine procurement: Bidders connected to one company only

This indicator is calculated by ensuring that all **Organization** only have one relation SUBMITS per **Bid** SUBMITTED_FOR a **Tender**.

3.5 Formal Verification of the Data Model

Formal models are high-level and abstract definitions meant to represent a wide variety of specific concepts. For this reason, they have been used extensively to introduce strict specifications of multiple concepts that otherwise sit in a fuzzy position where the very core elements are not clear enough. The main advantages that a formal model provides do not reside only in the fact that they can be as precise as needed, but also that they can be processed using formal and mathematical methods. The creation of a supporting formal model for, in this case, a data model, would lead to ensuring that its behaviour and structure are well defined, and thus, we could say that the data model has been formally verified.

Formal verification, as mentioned above, makes sure that the underlying concept behaves as it should be expected, but not only that. Depending on what formal model has been defined, the metaproperties of this can be easily transferred to the data model itself. That means that if a formal model is complete, sound, and decidable, so will the data model. In particular, completeness and soundness are tied to the ability to extract inferences from the data model; that is, any semantic reasoning that happens in and with the data captured by the data model will be valid, despite any possible setbacks. The decidability result means that any reasoning operation happening on or over the data model will be directed by an algorithm that will reach completion in a finite number of steps, thus avoiding infinite loops.

As stated above, the data model defined for CEDAR MVP can be formally verified by ensuring that it checks under a formal model. In particular, we will ensure that the data model complies with the formal model developed for the modal system **T** (Gabbay, 2001). This formal verification follows the idea behind (Blanco, 2024), where the topology, in this case of the data model, can be interpreted as an instance of relational semantics. The main difference lies in the use of Kripke-style semantics rather than the Routley-Meyer semantics used in the *opus citatis*.

The Kripke-style formal model for the system **T** is defined as a structure such that $M = <G, K, R>$. K is a non-empty set such that $K = \{\mathbf{a}, \mathbf{b}, \mathbf{c}, ...\}$, where **a**, **b**, **c** are non-empty theories comprised of an undetermined number of well-formed formulae. G is an element of K. And R is a reflexive relation defined over K. With all the above, this model is complete, sound, and decidable in the past (Chagrov, 1997; Kripke, 1959).

Given the data model defined for the MVP, we can easily codify it as follows. $|K| = 5$, with $K = \{\mathbf{a}, \mathbf{G}, \mathbf{c}, \mathbf{d}, \mathbf{e}\}$, where **a** = Public_Entity, **G** = Tender, **c** = Bid, **d** = Item, **e** = Organization. The cardinality of the elements of K is $|\mathbf{a}| = 5$, $|\mathbf{G}| = 7$, $|\mathbf{c}| = 6$, $|\mathbf{d}| = 7$, $|\mathbf{e}| = 5$. The different elements of **a**, **b**, **c**, **d,** and **e** are the attributes of each entity of the data model. In the case of the data model, **G** has been decided to be equal to Tender since it is the entity with the highest number of relations. Finally, the data model also has the following relations: $R\mathbf{aG}$, $R\mathbf{Ga}$, $R\mathbf{cG}$, $R\mathbf{Gd}$, and $R\mathbf{ec}$.

All this comes to show that the data model is, indeed, an instance of a Kripke-style model **M** as defined before. This leads to ensuring that any inferences that are made through the relations, or the usual rules of inference included in the system **T**, are valid, and the data inferred will hold. It also ensures that these inferences can be made, despite any possible growth in size. Finally, it ensures that there is an algorithm that provides a resolution for any inferences made within the model.

4 Data Protection and (Pseudo)Anonymisation

This chapter explores how the CEDAR project addresses the GDPR's data protection obligations, particularly focusing on pseudonymisation and anonymisation. It examines the legal, technical, and practical implications of handling sensitive data within a research context, highlighting how CEDAR implements de-identification strategies to ensure compliance and data security.

4.1 Key Definitions: Personal Data, Anonymised Data, and Pseudonymised Data

The GDPR establishes the main legal framework for processing personal data within the European Union. In the context of a research project, such as CEDAR, which involves large-scale data processing, a precise understanding of key terms is essential for ensuring legal compliance and safeguarding data subjects' rights.

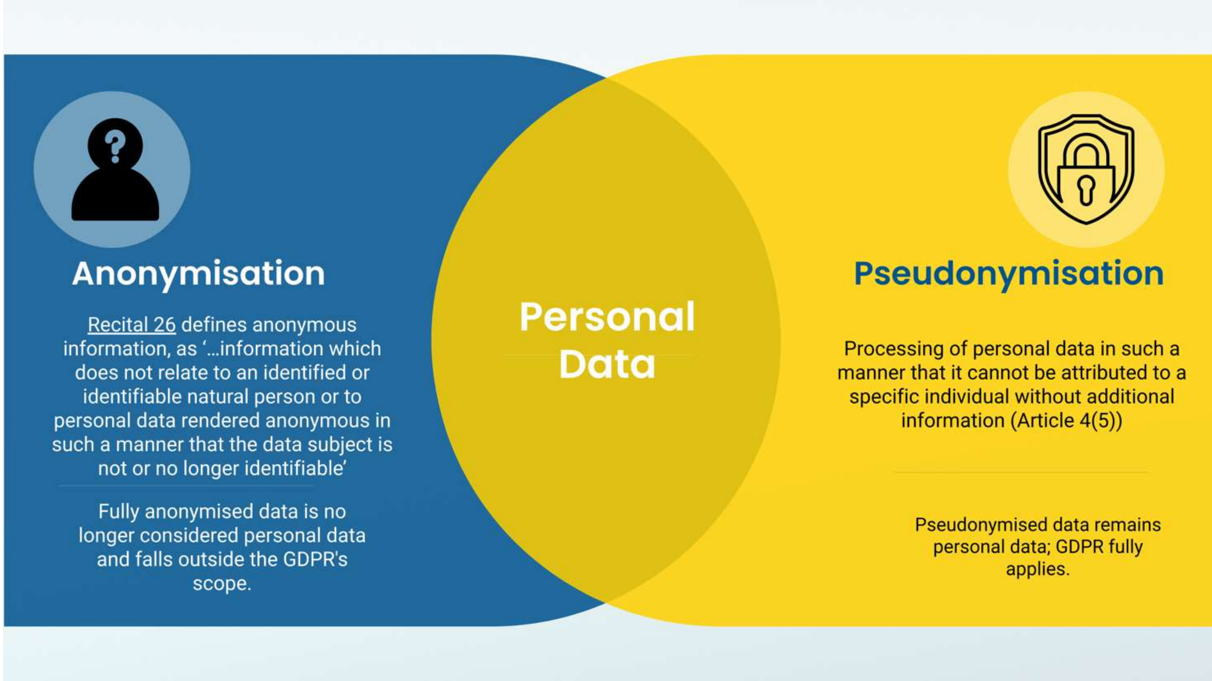
The first term that needs clarification is personal data. According to Article 4(1) of the GDPR, personal data is defined as "any information relating to an identified or identifiable natural person ('data subject'). An identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier, or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person." The concept of being "identified, directly or indirectly" is further explained in Recital 26 of the GDPR, which states: "To determine whether a natural person is identifiable, account should be taken of all means reasonably likely to be used, such as singling out the person, either by the controller or by another person, to identify them directly or indirectly." In simpler terms, personal data refers to any information that can be linked to an individual, allowing them to be identified. A person is considered "identifiable" if they can be recognized, either directly (e.g., by their name) or indirectly (e.g., through location data or online identifiers), using specific identifiers or characteristics. This means that personal data isn't limited to information like names, it can also include, in consideration of the context, the nature of the processing activity and the knowledge and means of the entity processing the data, other data that, when combined or analysed, can lead to identifying a person.

Still, within the scope of GDPR, pseudonymised data is defined in Article 4(5) as data that has been processed in such a way that it "can no longer be attributed to a specific data subject without the use of additional information," provided that this additional information is kept separately and protected through technical and organisational measures. Commonly, this involves replacing direct identifiers with codes or pseudonyms while securely storing the linkage key elsewhere. Importantly, pseudonymised data is understood as a security measure for personal data, for which GDPR still applies, as re-identification is still possible.

Anonymisation, in contrast, aims to eliminate any means of re-identification so that the data can no longer be linked to an individual. Properly anonymised data fall outside the scope of the application of GDPR, as Recital 26 clarifies that "personal data rendered anonymous in such a manner that the data subject is… no longer identifiable". Data can be simplified to a generalised level (aggregated) or transformed into statistics to ensure that individuals cannot be identified from it.

The measures taken to prevent identification must be permanent, making it impossible to revert the data to an identifiable form using additional information. Anonymisation must consider all reasonably viable methods for converting the data back to an identifiable form. Factors such as the cost of identification, the time required to identify the data subjects and the available technologies must be evaluated when assessing the likelihood of identification. Additionally, the controller should prepare for the possibility that, over time, advancements in technology could undermine the effectiveness of the anonymisation. In practice, the risk of re-identification can always persist, especially when datasets contain indirect identifiers or when combined with other data sources, making it difficult to achieve absolute anonymisation. Lastly, the anonymisation process constitutes data processing, meaning GDPR obligations remain applicable during that stage.

4.2 Anonymisation vs Pseudonymisation



Anonymisation

Recital 26 defines anonymous information, as '…information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable'

Fully anonymised data is no longer considered personal data and falls outside the GDPR's scope.

Personal Data

Pseudonymisation

Processing of personal data in such a manner that it cannot be attributed to a specific individual without additional information (Article 4(5))

Pseudonymised data remains personal data; GDPR fully applies.

Figure 13: Anonymisation vs Pseudonymisation.

While pseudonymisation and anonymisation are conceptually distinct under the GDPR, their practical application often reveals a more nuanced reality. Understanding the distinction between anonymisation and pseudonymisation can provide valuable guidance in determining the appropriate legal and technical frameworks for data processing within research initiatives. Indeed, anonymisation and pseudonymisation serve similar objectives, protecting data subject identities, but they differ fundamentally in outcome and legal implications. The distinction between the two is not always clear-cut in practice. European data protection authorities and courts have provided varying interpretations, sometimes viewing data anonymised according to a specific subject (relative anonymisation), while others apply a stricter, absolute standard.

The absolute anonymisation criterion reflects the traditional understanding of anonymisation, viewing it as a binary outcome: data is either completely anonymous or not at all. According to this perspective, data can only be considered anonymised if it is impossible for any party to re-identify individuals. Therefore, the standard for achieving anonymisation is high. The European Data Protection Board supports this stricter, more risk-averse interpretation. It maintains that anonymisation must remove all potential risks for re-identification, even by actors with significant resources or access to external datasets. While this approach offers strong privacy guarantees, it can pose practical challenges for research projects that require flexibility, particularly when some data linkage is necessary for longitudinal analysis or validation purposes.

Recently, jurisprudence and academic discussions have suggested a more nuanced approach that recognises whether data is truly anonymous depends on context. The capacity of the recipient party to reidentify the data differs from the divulging party's capacity. This element should be considered, as previously underlined by the CJEU in the Breyer Judgment (C-582/14) and the Scania Judgment (C-319/22). Recently, the CJEU has returned to the topic in the SRB v. EDPS (General Court of 2023, Case T-557/20), the Court held that data transmitted in the pseudonymised form to a third party did not count as personal data if the recipient had no "reasonable means" to re-identify the individuals. The judgment emphasised that determining personal data must consider the "recipient's perspective". The pseudonymised data may be considered effectively anonymous if the recipient has no additional information, or legal right to obtain it, or if doing so would be unreasonable. The Court noted that what matters is whether the recipient can re-identify the person, not merely whether the original controller's data is theoretically linkable. In the EDPS v. SRB case, the Advocate

General similarly opined (February 2025) that pseudonymised data shared with a third party may fall outside the definition of personal data if the risk of re-identification is "non-existent or insignificant".

These views reflect a line of reasoning dating back to the Breyer ruling, which held that identifiability is a relative concept depending on what means are reasonably likely to be used.

These developments highlight a divergence in interpretation. The EDPB and many national regulators maintain a conservative stance that if re-identification is possible in theory (even if only by the controller with the key), GDPR continues to apply. By contrast, the recent CJEU case law suggests that pseudonymised data might be considered anonymous when the recipient does not have reasonable means to re-identify the data subject.

For CEDAR, this divergence highlights the necessity for caution. It is recommended that the project treats all pseudonymised data as personal whenever the key exists and implements full GDPR safeguards accordingly. It's advisable to consider data anonymised only when it has been definitively proven that it is impossible to retrieve the identity of the data subject. This conservative approach ensures compliance regardless of legal uncertainty: all pseudonymised datasets remain under GDPR rules, triggering transparency, lawful basis documentation, and security obligations even during inter-consortium data exchanges.

4.3 Implementation of Pseudonymisation in CEDAR

Under the GDPR, controllers and processors must implement "appropriate technical and organisational measures" to secure personal data against risk. Furthermore, Article 25 explicitly names pseudonymisation as an example of a technique designed to implement data protection principles, such as data minimisation. The European Data Protection Board notes that pseudonymisation used "by design" can significantly reduce confidentiality risks and check 'function creep' by preventing unauthorised linking. Indeed, reducing the identifiability of the data subjects ensures that the amount of personal data processed is minimised to what is strictly necessary. Moreover, pseudonymisation is one of the measures that the controller and processor shall implement by Article 32 for the security of processing.

These provisions underscore that GDPR compliance for CEDAR requires proactive de-identification strategies to reduce privacy risks. Indeed, the system integrates pseudonymisation and secure data pipelines as a core safeguard. Before analysing the data, CEDAR stores raw data on encrypted servers and removes or masks identifiers, enabling the processing of only pseudonymised data. For example, after each pilot's data collection, identifying details are pseudonymised unless explicit consent dictates otherwise. CEDAR's Data Management Plan (D.7.2 published on 01.07.2024) consistently references these rules and specifies that only minimal, necessary personal data are collected and that analytic outputs use de-identified or pseudonymised inputs (with all processing logged and controlled).

Any gathered data will be securely handled throughout the entire duration of CEDAR to protect it from loss and unauthorised access. Before any data can be used, personal data will be pseudonymised, stored securely by the data controller, and only accessible to those authorised. Indeed, CEDAR adopts privacy-by-design and security-by-design principles. This approach ensures that personal identifiers are replaced with pseudonyms at the earliest possible stage in the data processing pipeline, thereby significantly reducing the risk of re-identification. By embedding privacy and security considerations into each phase of system development and data handling—from data collection and storage to analysis and dissemination—this methodology minimises exposure of sensitive information and aligns with ethical and legal standards for data protection. The systematic application of these principles ensures that individual identities remain obscured throughout the research process, thereby safeguarding participants' confidentiality.

More specifically, as described in the above-mentioned Data Management Plan, the initial phase of this process entails data preprocessing, wherein attributes are systematically classified into four categories: Identifiers (I), Quasi-identifiers (QI), Sensitive (S), and Non-sensitive attributes. Identifiers, such as full names and social security numbers, unambiguously disclose individual identities and must be expunged or substituted with pseudonyms to ensure compliance with data protection principles. Quasi-identifiers, including attributes such as date of birth, gender, geographic location, and occupation, do not individually reveal identity but may do so when combined with other data points. Sensitive attributes encompass highly confidential personal information, including but not limited to religious beliefs, sexual orientation, and political affiliations. In contrast, non-sensitive attributes do not contain personally or contextually sensitive content.

In the subsequent phase, a set of privacy-enhancing techniques will be systematically applied. These techniques include global recoding, local recoding, top-and-bottom coding, noise addition, microaggregation, PRAM, angelisation, and slicing. Each method will be employed to generate modified datasets that mitigate the risk of re-identification. The datasets will then be evaluated across two principal dimensions: disclosure risk and data utility for predictive modelling. Disclosure risk will be quantified using established privacy metrics, including k-anonymity, l-diversity, t-closeness, β-likeness, and distance-based risk measures, to ensure adherence to legal and ethical data minimisation and proportionality standards.

For datasets intended for data mining and machine learning tasks, the utility of the perturbed data will be evaluated in terms of data mining/ machine learning workloads. Predictive models will be built from the perturbed data, and the prediction accuracy will be used to assess the usefulness of these datasets. Typical measures to evaluate predictive performance in classification include Precision and Recall, (balanced) Accuracy, (weighted) F-score, Geometric Mean, and (weighted) AUC. Classification algorithms that will be considered include Random Forest, Bagging, XGBoost, and Support Vector Machine. Evaluation will be based on Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) for regression tasks, employing models such as linear regression, logistic regression, and regression trees. Clustering analyses will be assessed using cluster validity indices such as the Rand Index, Davies-Bouldin Index, Fowlkes–Mallows Index, and Silhouette Score.

Lastly, pseudonymisation is particularly valuable for the CEDAR project, which manages large-scale, multi-jurisdictional datasets to improve transparency and accountability in public governance. Within a consortium composed of partners from various EU Member States and non-EU, pseudonymisation functions not only as a privacy-enhancing technique but also as a legal tool that facilitates compliant data sharing. By replacing direct identifiers with coded values while retaining the possibility of re-identification through strictly controlled additional information, pseudonymisation enables CEDAR partners to exchange data across borders without disclosing personal identities, thereby significantly reducing legal and ethical risks under the GDPR. Implementing pseudonymisation for CEDAR is of particular importance. In the recent Guidelines on Pseudonymisation issued by the European Data Protection Board, pseudonymisation has been put forward as an effective additional measure to protect personal data during international transfers.

As previously outlined, the CEDAR project implements pseudonymisation at the point of data collection. By applying this measure at such an early stage, all personal data is de-identified before entering the shared processing environment, thereby significantly enhancing the security of data handling and exchange among consortium partners. The project's governance framework further supports this approach. The processing of personal data within the consortium is structured according to the Grant Agreement, the Description of Actions (DoA), and a shared Data Management Plan (DMP). The Data Management Plan outlines which datasets (from WP1–WP5) may be shared and under what conditions. Personal data is shared internally only in pseudonymised form and only with partners with legitimate project needs. After each pilot, collected personal data are encrypted/locked on secure servers, then processed or pseudonymised before any broader use. These safeguards also extend to external dissemination: public-facing resources such as the CEDAR website or open-access reports contain no direct identifiers, and any example data or outputs are vetted to avoid unintended re-identification. Furthermore, the Consortium agreement prohibits any partner from attempting to link codes back to individuals.

To further increase data protection and minimise re-identification risks, CEDAR employs a separation of duties strategy: one institution maintains the pseudonymisation key, while another holds the pseudonymised dataset. This division ensures that no single entity, aside from the original data source, can access both the identifiers and the corresponding data. Such an arrangement is by GDPR Article 32, which mandates the implementation of appropriate technical and organisational measures to ensure data confidentiality and resilience. The combination of strong encryption, physical isolation of the key, and stringent access controls ensures that, even in the unlikely event of a data breach, re-identification remains highly improbable.

4.4 Description of the Pseudo-Anonymisation Tool for the MVP

At the current stage of the CEDAR project, all the datasets collected by the pilot users requiring pseudonymization are tabular. For this reason, the MVP version of the pseudonymization tool is designed around the structured format of tabular data.

Among the privacy threats commonly considered when anonymising structured data (see D2.1 Section 4.4.1), **Identity Disclosure** (where an attacker can link a record in the released data to a specific individual) **was found to be the most prominent risk in all the MVP use cases.** Concretely, the problem consists of preventing an identity present in the tabular data from being re-identified by an attacker via singling out. For this reason, k-anonymity was selected as the appropriate risk disclosure measure.

Concrete example:

The tool intends to protect the identity of the clerk responsible for the Tender bureaucracy. The clerk's name is hashed, but an attacker could still re-identify the clerk if, for example, it is known that the clerk is the only person working in a specific geographical area represented in the tabular data. By generalizing the location attribute during pseudonymization, it becomes significantly more difficult for the attacker to infer the clerk's identity.

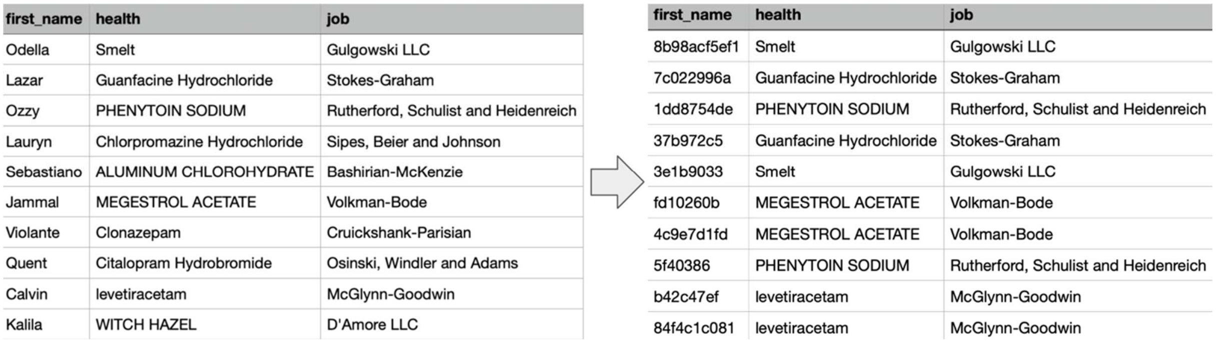
To improve the k-anonymity of tabular data, Global Recoding (generalisation) and Local Recoding were selected as the main approaches to implement.

Global recoding (a.k.a. generalisation) combines several categories to create more general categories. The application of global recoding on a categorical attribute *V,* results in a new attribute *V'* with fewer possible values. For a continuous attribute, *V* is replaced by a discretised version of *V*. The main objective of global recoding is to divide the tuples in the dataset into a set of disjoint equivalence classes and then transform the attribute values of the tuples in each equivalence class to the same format. Currently, the tool handles the generalisation of several attribute types in the following manner:

- Date generalisation: a timestamp is successively generalised by removing seconds, then minutes, then hours, etc.
- Location generalisation: due to the current absence of a specific location format in CEDAR pilot data, the generalisation is a general string generalisation where the string is tokenised, and a frequency count of the tokens in the attribute column is computed. Only tokens appearing more than *k* times are kept.
- Amount generalisation: successive rounding of the amount value until k-anonymity is reached.

Local recoding recodes into broader categories when necessary. The replacement can be partial, i.e., only some occurrences of the attribute *V* are replaced. Therefore, it can have a lower cost in terms of information loss. The current implementation of local recoding in the Pseudonymization Tool transforms a tabular dataset into a k-anonymous version by recoding QI values belonging to equivalence classes of size < *k* with new values identical to those of the 'closest' datapoint (by attribute distance) until the new equivalence class size is greater than or equal to *k*.

Table 3: An example of Local Recoding.



first_name	health	job
Odella	Smelt	Gulgowski LLC
Lazar	Guanfacine Hydrochloride	Stokes-Graham
Ozzy	PHENYTOIN SODIUM	Rutherford, Schulist and Heidenreich
Lauryn	Chlorpromazine Hydrochloride	Sipes, Beier and Johnson
Sebastiano	ALUMINUM CHLOROHYDRATE	Bashirian-McKenzie
Jammal	MEGESTROL ACETATE	Volkman-Bode
Violante	Clonazepam	Cruickshank-Parisian
Quent	Citalopram Hydrobromide	Osinski, Windler and Adams
Calvin	levetiracetam	McGlynn-Goodwin
Kalila	WITCH HAZEL	D'Amore LLC

first_name	health	job
8b98acf5ef1	Smelt	Gulgowski LLC
7c022996a	Guanfacine Hydrochloride	Stokes-Graham
1dd8754de	PHENYTOIN SODIUM	Rutherford, Schulist and Heidenreich
37b972c5	Guanfacine Hydrochloride	Stokes-Graham
3e1b9033	Smelt	Gulgowski LLC
fd10260b	MEGESTROL ACETATE	Volkman-Bode
4c9e7d1fd	MEGESTROL ACETATE	Volkman-Bode
5f40386	PHENYTOIN SODIUM	Rutherford, Schulist and Heidenreich
b42c47ef	levetiracetam	McGlynn-Goodwin
84f4c1c081	levetiracetam	McGlynn-Goodwin

4.4.1 Pseudo-Anonymisation Tool Pipeline

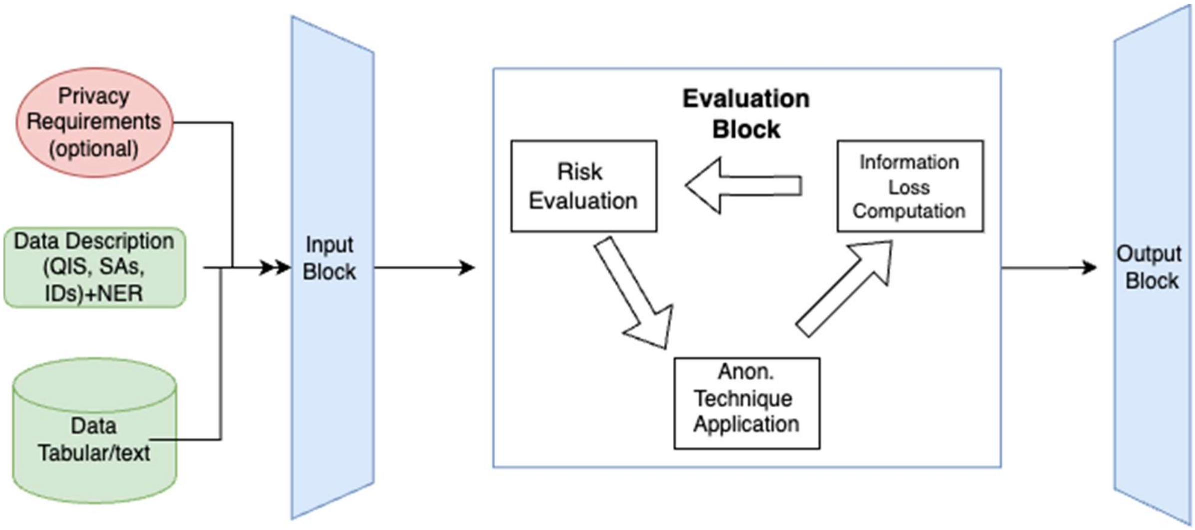

Privacy Requirements (optional)
Data Description (QIS, SAs, IDs)+NER
Data Tabular/text
Input Block
Evaluation Block
Risk Evaluation
Information Loss Computation
Anon. Technique Application
Output Block

Figure 14: A scheme of the Pseudo-anonymisation Tool Pipeline.

The functionalities of the tool can be described as a series of blocks working in succession to process the data. In its current state, the Pseudonymization tool includes the following components:

- **Input Block** essentially controls data integrity and correctness of the format of the following inputs:
  - **Data:** Tabular data to pseudonymize; supports single or multiple instances of .txt, .csv, and .xlsx files.
  - **Descriptor Information** (data model oriented):
    - List of quasi-identifier (QI) columns or entity types.
    - List of identifiers to hash.
    - Data type specification (optional): a CSV file where each row contains a column name (or entity type) and its data type (e.g., integer, categorical, text).
  - **Anonymisation Requirements:** Desired anonymisation level; by default, the tool targets 3-anonymity.
- **Evaluation Block** applies both pseudonymization techniques—global and local recoding—and selects the approach that successfully achieves k-anonymity while minimising information loss.
- **Output Block** mainly takes care of making available the following outputs:
  - A pseudonymized version of the input data, preserving the original format (CSV or JSON for structured data; text for unstructured data).
  - A pseudonymization report including:
    - An anonymisation quality score, which evaluates the trade-off between:
      - **Information Loss:** measured by Mutual Information (for categorical data) and Mean Variance (for numerical data).
      - **Disclosure Risk:** percentage of records that can be linked between the original and anonymised datasets.
      - **Predictive Performance** (optional): Accuracy (for classification tasks) and R^2 (for regression tasks).
    - A higher score indicates better data quality post-anonymisation.
  - A hashed entities dictionary, enabling recovery of original identifiers from the hashes.

Table 3 reports an example of input (left) and output (right) data. The sensitive attribute is *first_name*, and the QIs are *health* and *job*. The pseudo-anonymised version of the input data satisfies the 2-anonymity constraint.

4.4.2 Description of the APIs

To integrate with existing data processing pipelines, the Pseudonymization Tool is designed to operate via a REST API interface. This enables the tool to receive requests, access data, perform pseudonymisation, and deliver results in an automated and scalable way.

The process follows this general workflow:

1. **Request Handling:** A request is sent from the data processing pipeline to the Pseudonymization Tool via a REST API call (typically using GET or POST methods). This request includes:
   a. The anonymisation requirements (e.g., desired k-anonymity level, columns to pseudonymise).
   b. The address (URI) of the data storage bucket containing the tabular data to be pseudonymised.
2. **Processing:** Upon receiving the request, the tool retrieves the specified data from the indicated bucket, applies the pseudonymisation techniques (as described previously: global recoding, local recoding, and hashing of identifiers), and generates the corresponding outputs.
3. **Output Delivery:** Once processing is complete, the tool:
   a. Saves the pseudonymised dataset into a designated output bucket.
   b. Stores a hash dictionary in a separate bucket, allowing future reference or recovery of original identifiers if needed.
   c. Optionally generates and stores a pseudonymisation report summarising information loss, disclosure risk, and predictive performance metrics.

5 Conclusion

This deliverable demonstrates significant progress towards the CEDAR project's goal of fighting corruption in public procurement through the preparation of high-quality, analytics-ready datasets and the definition of indicators of lack of transparency.

The analysis of datasets collected until M18 has been crucial in understanding their utility for the three use cases and identifying gaps that necessitated additional data. In particular, to address the issue of data scarcity, we developed a synthetic data generation tool with a specific application for the Slovenian use case.

A key achievement has been the development of a data model designed to effectively represent the main entities involved in public procurement and their relationships. This model includes a common set of entities for the three use cases and a comprehensive set of attributes to support data analytics for corruption identification.

Two important aspects of this process have been data quality and protection. To address data quality, we studied how to guarantee data quality from a legal perspective throughout the data lifecycle (collection, storage, use, protection, archiving, and deletion), emphasising adequacy, relevance, proportionality, and accuracy. To address data protection, we developed a (pseudo) anonymisation tool to minimise the risk of identity disclosure, ensuring GDPR compliance.

In conclusion, the work carried out in these months has provided the necessary tools and frameworks for training and testing machine learning algorithms aimed at combating corruption in public procurement.

Looking ahead, the next steps include:

- Continuing to improve the data model to ensure it accurately represents the data collected so far and extending it to cover additional use cases and data domains.
- Implementing and testing the risk indicators defined over the data model.
- Thoroughly testing the pseudo-anonymisation tool to ensure its proper functioning and to identify any additional features or techniques (e.g., other data types or privacy-enhancing techniques) that may need to be included.

6 Bibliography

Blanco, J. M. (2024). A Formal Model for Reliable Data Acquisition and Control in Legacy Critical Infrastructures. *Electronics, 13*(7).

Chagrov, A. Z. (1997). *Modal Logic.* Oxford University Press.

Gabbay, D. M. (2001). *Handbook of Philosophical Logic.* Springer Netherlands.

Kripke, S. (1959). A Completeness Theorem in Modal Logic. *Journal of Symbolic Logic, 24*(1), 1-14.

Funded by the
European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101135577