Knowledge Graph based Representation to Extract Value from Open Government Data

An Application to the Public Procurement Domain

Kawtar YOUNSI DAHBI[1], Dalila CHIADMI[2], Hind LAMHARHAR[3]
Mohammed V University of Rabat
Rabat, Morocco[1, 2, 3]

***Abstract*—Open government data refers to data that is made available by government entities to be freely reused by anyone and for any purpose. The potential benefits of open government data are numerous and include increasing transparency and accountability, enhancing citizens' quality of life, and boosting innovation. However, realizing these benefits is not always straightforward, as the usage of this raw data often faces challenges related to its format, structure, and heterogeneity which hinder its processability and integration. In response to these challenges, we propose an approach to maximize the usage of open government data and achieve its potential benefits. This approach leverages knowledge graphs to extract value from open government data and drive the construction of a knowledge graph from structured, semi-structured, and non-structured formats. It involves the extraction, transformation, semantic enrichment, and integration of heterogeneous open government data sources into an integrated and semantically enhanced knowledge graph. Learning mechanisms and ontologies are used to efficiently construct the knowledge graph. We evaluate the effectiveness of the approach using real-world public procurement data and show that it can detect potential fraud such as favoritism**.

Keywords*—Knowledge graph; open government data; knowledge graph construction; public procurement; fraud detection

I. INTRODUCTION

Open Government Data (OGD) is a concept that continues to prosper and evolve. It includes all data collected or produced by public administrations that are made available for the public to be freely used [1]. The publication of government data has a significant impact that could be identified through multiple aspects: fostering innovation, improving transparency, public accountability, and collaboration, and improving citizens' quality of life [2] [3]. However, despite these efforts, technical barriers limit the effective use of OGD. For example, users may have difficulty finding relevant data due to the large volume of information and the diversity of portals [4][5]. Additionally, the data may be presented in unstructured formats, requiring manual transcription and significant time to process [2], [6], [1]. Furthermore, a single dataset may not be enough to fulfill user requests, so combining data from multiple sources is often necessary. However, syntactic, structural, and semantic heterogeneity can make it challenging to integrate this data effectively [7]. Therefore, using OGD in its raw form requires technical expertise and significant time and effort to discover, process, integrate, and analyze the data. Given the potential benefits of OGD and the technical barriers to its effective exploitation, exploring new approaches and technologies that can facilitate its use is crucial. One promising avenue for such exploration is the use of knowledge graphs, which have already been successfully implemented by tech giants such as Google and widely used in several contexts [8] – [12]. A knowledge graph is a powerful tool for organizing and analyzing complex information, such as that contained in OGD. It is a semantic graph that captures information about entities and their relationships in an easily machine-processable way [12]. The use of semantics in knowledge graphs allows for a more nuanced and contextually rich understanding of the data, leading to more accurate and insightful analyses. Knowledge graphs offer also a centralized solution for integrating different types of data from heterogeneous sources, providing end-users with a single point of access. Furthermore, knowledge graphs support multiple advanced applications, such as Q&A systems, semantic search, reasoning, and knowledge inference, enabling the development of smart services that extract value from OGD[13] [14].

In this paper, we propose a generic approach that explores the potential of knowledge graphs to transform open government data (OGD) into valuable knowledge. The approach involves constructing a knowledge graph from structured, semi-structured, and unstructured OGD data sources, and offers the following contributions:

- Representing OGD in a semantically rich and machine-processable format to enhance their usefulness.
- Integrating heterogeneous data sources into a centralized solution represented by a knowledge graph, which enables users to have a unified and comprehensive view of the data.

By adopting this approach, we aim to address the challenges related to data processability and integration and to provide users with easily usable data that can be effectively and efficiently utilized. The proposed approach offers a powerful solution for maximizing the value of OGD and encapsulates the technical difficulties associated with processing open government data (OGD). The rest of this paper is organized as follows: Section II presents related works, Section III gives an overview of the proposed approach to construct a knowledge graph from OGD, and Section IV presents the use case related to the public procurement domain, which aims to construct a

knowledge graph based on public procurement data and performs advanced analysis to detect anomalies such as favoritism and overpricing. The study concludes in Section V.

II. Related Works

In this section, we present related works that propose the construction of a knowledge graph to extract value from Open Government Data (OGD). We give a brief overview of each work and then draw their main limitations.

The authors in [15] propose the construction of a knowledge graph related to the cadaster in the Netherlands. The knowledge graph is exploited to offer improved data browsing, analysis for urban planning, and the development of location-aware chatbots. However, the approach considers only RDF-like datasets.

Authors in [16] propose the construction of a knowledge graph from Zaragoza's open data to provide a single point for the city's knowledge. Semantic enrichment and transformation are supported through the usage of scripts to transform datasets to RDF. However, the approach doesn't propose dataset integration.

In [17], the authors propose to construct a knowledge graph for the description of public services, the objective of the knowledge graph is to provide users with personalized information about different public services based on their profile and circumstances. The implementation of the knowledge graph was achieved by creating the schema defining entities, attributes, and relationships and populating the graph using GRAQL queries.

The TBFY project [18] [19] proposes the construction of a public procurement knowledge graph to support transparency. For semantic enrichment, the approach uses two ontologies. The transformation to RDF is supported by the RML Mapper tool which automates the generation of RDF triples based on the RML mapping. However, it requires the manual definition of the mapping between the data sources and the ontologies.

Authors in [20] propose the construction of a knowledge graph to support the supervision and analysis of fiscal projects funded by the European Union. The proposed approach collects data from the Open Data API provided by the Greek Ministry of Economy and Finance. Semantic enrichment is based on two ontologies. The approach proposes the usage of the ETL unified views for Data transformation to RDF and semantic enrichment and publication of data in an RDF triple store. Data can be retrieved via SPARQL queries. Performance indicators were defined to assess the state of the project and Density-Based Spatial Clustering of Applications with Noise, (DBSCAN) was used to identify Red Flags.

Authors in [21] propose a solution to support budget transparency. The proposed approach is based on the creation of a knowledge graph from data related to the public budget. Data is published by the budget office in the form of an XML file: an annual file for budget distribution and monthly files for monitoring budget execution. The data is semantically enriched with the National Budget ontology and transformed into RDF through the use of an ETL tool. For data exploitation, the approach proposes the publication of data through a SPARQL access point as well as a set of solutions for data visualization.

The related works presented to provide a background for extracting value from open government data (OGD) through their transformation into knowledge graphs. However, they have limitations that we present below. Firstly, most approaches are either domain-specific or related to specific data sources and cannot be adapted to diverse contexts and domains. Secondly, data transformation to RDF is carried out using scripts, ETL, or mapping languages. This process is tedious, time-consuming, and difficult to implement. It requires the intervention of the user to specify the mapping between data sets and semantic models, which requires a good understanding of the dataset content and structure, domain knowledge, and technical expertise. Thirdly, the majority of approaches focus on dataset transformation without proposing solutions for dataset integration, such as entity linking. When it is proposed, it is not generic and is done for specific types of entities based on linking rules specified by the users. Lastly, most approaches consider only structured data sources. This excludes a large proportion of OGD that is published in unstructured formats.

Therefore, in this work, we aim to address these challenges by proposing a generic and domain-independent approach to constructing a knowledge graph. The proposed approach considers structured, non-structured, and unstructured data sources, and provides an efficient solution to semantically transform OGD with a focus on data integration.

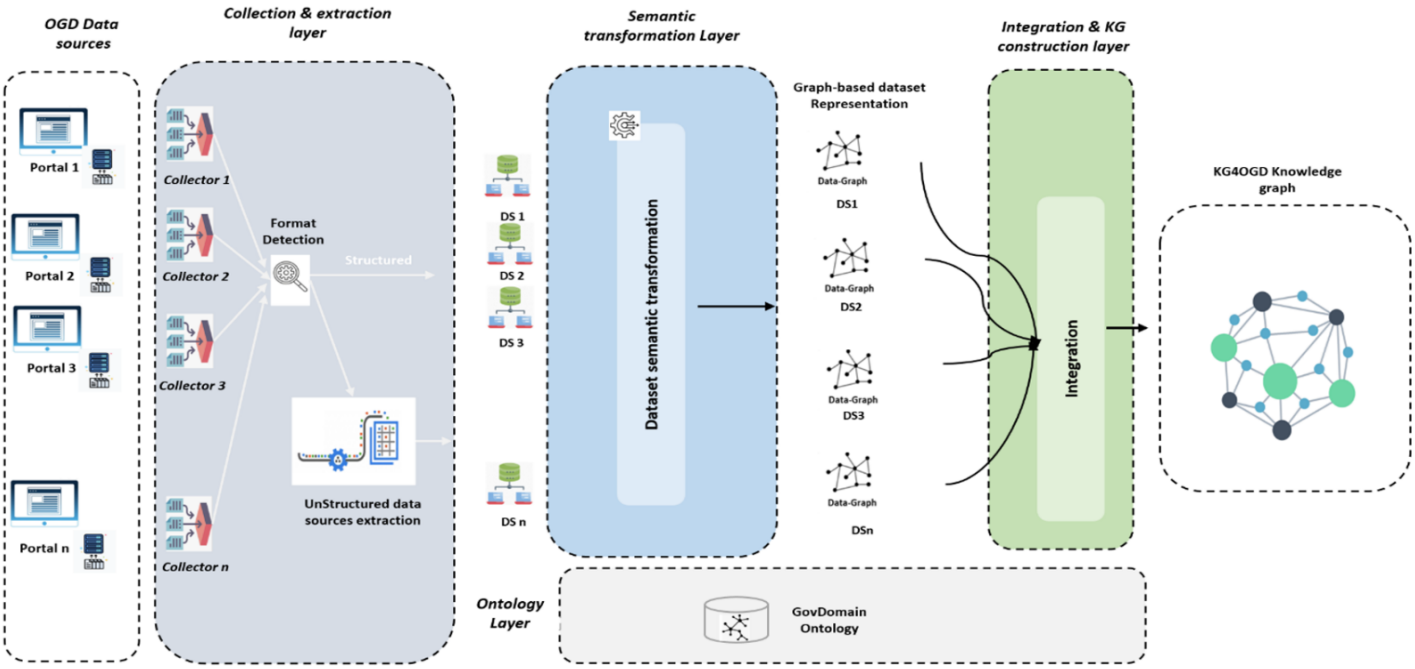

Fig. 1. Overview of the proposed architecture.

III. The Proposed Approach

This section presents a detailed description of the proposed approach for constructing a knowledge graph from Open Government Data sources. The approach is based on a layered architecture that can collect and extract data from structured, semi-structured, and unstructured heterogeneous government data sources. The collected data undergoes semantic enhancement and transformation, after which it is represented as a knowledge graph that offers an interconnected, integrated, and unified representation of data. The constructed knowledge graph can indeed be used through a plethora of applications, such as Q&A, semantic search, knowledge reasoning, and advanced analytics capabilities. These features enable to provide open government data users with useful insights and valuable information.

The proposed architecture depicted in Fig. 1 comprises three main layers: *The Collection and Extraction* layer, the *Semantic Transformation layer*, and the *Integration and Knowledge Graph Construction* layer. These layers are further enhanced by a domain ontology for the representation of domain knowledge. We present below a description of these layers.

A. Collection & Extraction Layer

The Collection and Extraction layer is responsible for discovering and collecting data sources published on various government portals (Pi). Algorithm 1 presents the algorithm that outlines the process performed by this layer.

To discover relevant data sources, the layer employs a set of collectors that use web scraping techniques and can be configured to filter data sources based on keywords, data format, or other specific criteria (CFi) (Line 6 Algorithm 1). This layer is designed to retrieve both structured (datasets) and unstructured data sources, including those published in non-dataset formats such as reports, PDFs, and databases (Line 7 Algorithm 1).

The layer performs format detection (Line 8 Algorithm 1), and structured sources are directly transferred to the Semantic Transformation Layer (Line 10 Algorithm 1).

For unstructured data sources, the layer performs additional processing to extract structured data. For instance, it uses optical character recognition (OCR) and natural language processing (NLP) techniques to identify and extract tables from unstructured data sources, such as PDFs (Line 11 Algorithm 1). Each extracted table is considered a dataset and is transferred to the Semantic Transformation Layer for further processing (Line 14 Algorithm 1).

Algorithm 1: Collect & extract datasets

```
Input: - P_i government data portal
       - CF_i Filter criteria
Output: D = {datasets}
S← {}
D← {}
DiscoverDataSource(P_i , CF_i)
S←RetrieveDataSources()
For each s_i in S
    fs_i←DetectFormat(s_i)
    If fs_i is structured then D.add(fs_i)
    Else IdentifyTable(s_i)
        For j← 1 to n
            t_ij← ExtractTable (s_i)
            D.add(t_ij)
        EndFor
    EndIf
EndFor
Return D
```

B. Semantic Transformation Layer

The Semantic Transformation Layer is responsible for adding semantics to the data and transforming it into a machine-processable and understandable format. This layer aims to enhance the usefulness of the collected and extracted datasets by enriching them with semantic information. For this purpose, the layer uses the GovDomain ontology, a domain ontology that formalizes and represents domain knowledge. The Gov domain ontology is intended to model a consensus among governments on the concepts and relationships that exist in published datasets. It is thus an essential tool to solve the problem of semantic heterogeneity.

The semantic transformation layer takes as input the datasets collected and extracted by the Collection & Extraction layer. For each dataset di, it generates an RDF graph-based representation of the dataset GD-di, which is semantically enriched with the GovDomain ontology. The process performed by this layer involves two main steps depicted in Fig. 2: Semantic model construction and RDF graph generation.

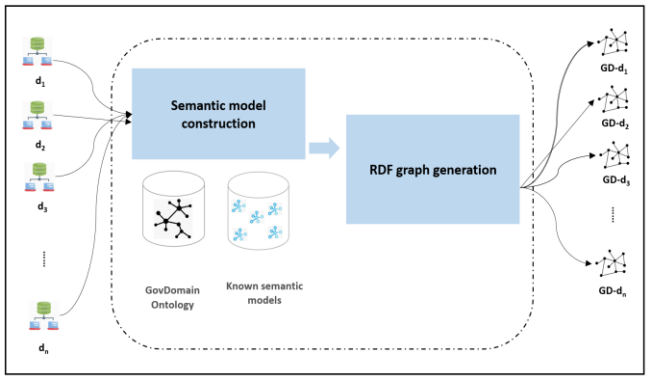

Fig. 2. Semantic transformation process.

The first step consists of constructing a semantic model for the dataset. This model is a requirement for semantic enrichment. It aims to define the structure of the dataset in terms of concepts and relations from the GovDomain ontology. Specifically, it establishes the mapping between the attributes of the dataset and the concepts or properties of the ontology, as well as the relationships between these attributes in terms of ontology relations.

For example, Fig. 3 depicts the semantic model of a public procurement dataset, which includes information on public contracts awarded by public entities, their details (title, reference, description, price, and duration, and the purchasing organization and selected supplier.

Constructing the semantic model for a dataset involves two sub-steps. The first sub-step involves mapping the dataset's attributes to their corresponding semantic types following the domain ontology. A semantic type can be a class URI or a combination of a class and a property from the GovDomain ontology. For instance, in the public procurement dataset, the attributes "Acheteur Public", "Fournisseur "," and "Titre" are mapped to their respective semantic types, which are (PCO: organization, orgName), (PCO: supplier, SupplierName), and (PCO: contract, title)). The second sub-step aims to establish the relations between the attributes in terms of ontology relations. For instance, the relations (PCO: Hasbuyer) and (PCO: HasSupplier) link the class (PCO: Contract) respectively with the classes (PCO: Organization) and (PCO: Supplier).

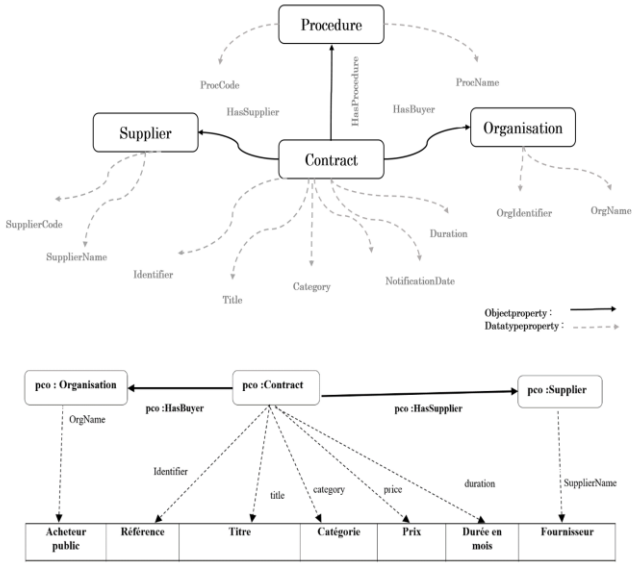

Fig. 3. Example of a semantic model for a dataset related to public procurement.

The manual creation of a semantic model for a dataset is a challenging and time-consuming task, which demands an extensive understanding of the dataset's content and structure, as well as domain-specific knowledge and technical expertise. To address this challenge, we propose an automatic approach to construct the semantic model of a dataset, thereby removing the need for manual intervention in the process. The proposed approach utilizes machine learning to automatically learn the dataset's semantic model by constructing it based on a set of known semantic models that serve as the training data. We employ the algorithm proposed by [22], which offers a method to learn semantic models of structured data sources by mapping them to a domain ontology. The algorithm constructs a weighted graph that represents the space of potential semantic models for a dataset, and provided a ranked list of the top-k best semantic models.

Based on the constructed semantic model the second step (Fig. 3) involves the RDF graph generation. The resulting model is formalized using the RDF Mapping Language (RML). RML [23] allows for the definition and expression of mapping rules between data sources and the RDF model, describing how the existing data can be represented according to the RDF model. The RML mapping document is interpreted by an RML processing engine and used to automatically generate RDF triples from the dataset. The resulting RDF graph provides a graph-based representation of the dataset, thereby enabling machine-readable and machine-understandable access to the dataset's content.

C. Integration and Knowledge Graph Construction Layer

The Integration & Knowledge Graph Construction Layer is responsible for the integration of data into a knowledge graph. This layer takes as input the RDF graphs that result from the

semantic transformation of datasets. The integration process involves two steps, which are outlined in Algorithm 2

Algorithm 2: KG construction & integration

```
Input: -GD-d1, GD-d2…. GD-dn: RDF graphs related to datasets { d1, d2…. dn }
       -C: List of ontology classes described in datasets { d1, d2….dn }
Output: KG4OGD knowledge graph.
KG4OGD ← ∅
For each GD-di
    KG4OGD ← KG4OGD.Add(GD-di)
EndFor
For each cp from C
    Bp ← ConstructBloc(cp)
        For each (ei,ej) ∈ Bp
            If Comp (vi,vj)= true then
            t←ConstructIdentityTriple(ei,ej)
            KG4OGD ← KG4OGD.Add(t)
            EndIf
        Endfor
EndFor
Return KG4OGD
```

The first step involves publishing and consolidating all datasets into the knowledge graph (Line 8 Algorithm 2). The second step is crucial for completing data integration and involves entity alignment. Entity alignment refers to the process of identifying and mapping coreferent entities, which are entities that refer to the same real-world entity but with different URIs [24]. This step involves a pairwise comparison of all entities that have been published in the graph. To improve efficiency, we use a blocking technique [25] that reduces the number of comparisons and identifies potential candidates for entity alignment. We apply a class-based partitioning approach to construct blocks, which groups potential co-referenced entities into blocks based on their class (Line 11 Algorithm 2). This method reduces the quadratic complexity of the process by limiting the number of comparisons required. After constructing the blocks, we compare the entities belonging to the same block by examining their discriminative properties, which are unique properties that identify entities of a particular class (Line 13, Algorithm 2). These discriminative properties are described in the domain ontology by adding owl:hasKey axioms. If two entities have equal values for these properties, we consider the entities to be co-referent, and we add identity links between them to the knowledge graph using the owl:sameAs property (Line 14 &15, Algorithm 2).

D. Advantages of the Proposed Approach

The proposed approach makes significant contributions to the field of Open Government Data (OGD). Firstly, it enables the representation of OGD in a structured and semantically rich format that can be easily processed and interpreted by machines. This overcomes the limitations of existing data formats and promotes greater interoperability and reuse of OGD. Secondly, the approach provides an integrated representation of OGD that allows users to view all published data in a unified and complete manner. By utilizing knowledge graphs to integrate data from multiple OGD sources, the proposed approach facilitates data discovery and analysis and enables the identification of previously unseen patterns and relationships across diverse datasets.

When compared to related works, the proposed approach offers significant contributions. Firstly, it addresses the challenge of integrating diverse data sources with structured, semi-structured, or unstructured formats. The approach automates the process of semantically transforming datasets, which can be time-consuming and error-prone when done manually. This is achieved by utilizing learning mechanisms and mapping languages. Furthermore, the proposed approach is domain-independent and presents a generic solution to integrate heterogeneous datasets. It accomplishes this through a schema alignment and entity alignment approach, which is designed to handle all types of entities without the need for human intervention to specify linking rules.

IV. USE CASE STUDY

In this section, we present a use case that implements the proposed architecture and demonstrates its operationalization. The use case focuses on the public procurement domain, which is a critical government sector for ensuring transparency and accountability in government spending.

A. Context

Public procurement is a significant part of government spending, but it is also a domain prone to corruption and fraud [26] [27] [29]. Public Procurement (PP) is a distinct area within Open Government Data (OGD) that plays a critical role in ensuring transparency and accountability in government spending. By releasing public procurement data as open and accessible, governments and civil society organizations can promote fair competition, identify patterns of corruption and fraud, and hold public officials accountable.

The use case aims to implement the proposed architecture in the context of public procurement in France. By integrating public procurement data from several open government data sources into a knowledge graph, we can apply anomaly detection algorithms to identify instances of fraud such as favoritism and overpricing. Favoritism in public procurement [28] refers to the practice of giving preferential treatment to a particular supplier, contractor, or bidder. Overpricing [29], on the other hand, refers to the practice of charging prices that are higher than what would be considered reasonable or fair.

B. Data Sources

In the French context, there exists a multitude of public procurement data sources, but for our specific purposes, we have opted to primarily utilize three sources: the BOAMP, Essential Data on Public Procurement (DECP), and the SIRET database (Cf. Table I).

[1]BOAMP is a French government platform that publishes public procurement notices and announcements, as well as detailed information about awarded contracts. The platform also provides an API that allows users to retrieve data in both JSON and XML formats

[1] https://www.boamp.fr/

Essential Data on Public Procurement (DECP) is a database that features structured and standardized information regarding all public procurement procedures. This expansive dataset includes details such as the contracting authority's name and location, the type and value of the contract, and the selected supplier. These data are consolidated in the French open government data portal[2].

The SIRENE[3] the database is a comprehensive registry of all business entities operating in France, providing information regarding each entity's legal status, activity sector, and location. By leveraging SIRENE data, we can identify the entities involved in public procurement procedures and establish links with other pertinent datasets.

TABLE I. PUBLIC PROCUREMENT DATA SOURCES

Data sources	Publisher	Format	Frequency of update
BOAMP	Direction of Legal and Administrative Information (DILA).	XML,JSON	Daily
Essential Data on Public Procurement (DECP)	The Ministry of the Economy, Finance, and Industrial and Digital Sovereignty	XML , JSON	Monthly
SIRENE Database	French National Institute of Statistics and Economic Studies (INSEE)	CSV	Quartly

By integrating and linking these various data sources in a knowledge graph, we can create a comprehensive and well-organized representation of public procurement data. To achieve this, we need to implement a domain ontology that captures and formalizes knowledge relevant to the field of public procurement

C. Ontology Development

Developing a domain ontology is a requirement for constructing the knowledge graph as it plays a role in the semantic enrichment of government data sources. To this end, we constructed a domain ontology for the public procurement domain by reusing the PCO ontology [30].

The PCO ontology, known as the Public Contracts Ontology, is a domain-specific ontology that models the essential concepts and relationships within the public procurement domain. Its primary objective is to support the integration and analysis of public procurement data from various sources. The ontology encompasses a broad range of concepts and relationships that apply to public procurement, such as procurement notices, contracts, suppliers, products and services, and tendering procedures. Additionally, it includes concepts linked to the legal and regulatory frameworks that oversee public procurement, such as procurement rules and regulations, procurement authorities, and procurement methods.

For the implementation, we have customized the PCO ontology to match the French context. This involved correlating the concepts and relationships in the ontology with the appropriate terms and structures in the French public procurement domain. This adaptation (Fig. 4) ensures that the constructed knowledge graph represents the relevant aspects of the French public procurement domain accurately and is tailored to our use case's particular needs.

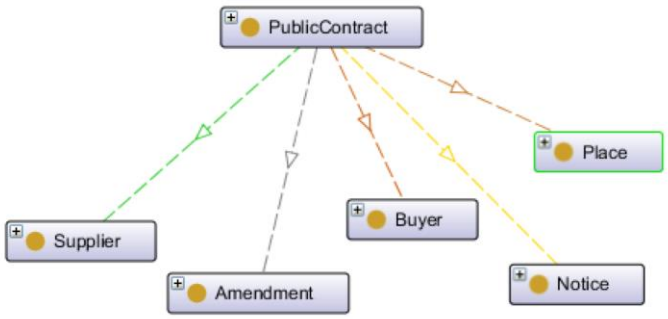

Fig. 4. Public procurement domain ontology implemented in protégé tool

D. Knowledge Graph Construction

The initial step in constructing the knowledge graph involved collecting data from identified data sources. We automated the data collection process by developing data collection scripts using Python and leveraging APIs. We utilized various APIs, such as the BOAMP API, and the API offered by the French government data portal, to collect structured data in different formats, including XML, JSON, and CSV. Fig. 5 shows an example of the collected data.

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

[3] https://www.data.gouv.fr/fr/datasets/base-sirene-des-entreprises-et-de-leurs-etablissements-siren-siret/

```
-<MANAGEMENT>
-<REFERENCE>
<IDWEB>22-160170</IDWEB>
<HTML_NAME>22-160170.html</HTML_NAME>
-<INDEXING>
<PUBLICATION_DATE>2023-01-04</PUBLICATION_DATE>
<END_OF_DISSEMINATION_DATE>2023-02-02</END_OF_DISSEMINATION_DATE>
-<DESCRIPTORS>
-<DESCRIPTOR>
<CODE>274</CODE>
<TITLE>Services</TITLE>
</DESCRIPTOR>
</DESCRIPTORS>
<PUBLICATION_DEPARTMENT>33</PUBLICATION_DEPARTMENT>
<OBJECT_SUMMARY>Tender for coordination of health and safety measures </OBJECT_SUMMARY>
</INDEXING>
</MANAGEMENT>
-<DATA>
-<IDENTITY>
<DENOMINATION>COBAS represented by SODEREC</DENOMINATION>
<ADJUDICATOR_NUTS>FRI12</ADJUDICATOR_NUTS>
<ADDRESS>31 Armagnac Street</ADDRESS>
<POSTCODE>33088</POSTCODE>
<CITY>Bordeaux Cedex</CITY>
<EMAIL>ssonrel@lasoderec.com</EMAIL>
<URL>http://www.lasoderec.com</URL>
<BUYER_PROFILE_URL>http://www.marches-securises.fr</BUYER_PROFILE_URL>
<PARTICIPATION_URL>http://www.marches-securises.fr</PARTICIPATION_URL>
<DOCUMENT_URL>http://www.marches-securises.fr</DOCUMENT_URL>
```


a-Example of a dataset collected using the BOAMP API


```
<PublicContract>
  <object>Renovation of the municipal stadium lighting</object>
  <uid>20006340200016201 9B040100</uid>
  <place>
    <name>ANNECY</name>
    <code>74000</code>
    <typeCode>Zip Code</typeCode>
  </place>
  <contractType>Framework agreement</contractType>
  <dataPublicationDate>2019-08-21+02:00</DataPublicationDate>
  <Buyer>
    <name>ANNECY CITY COUNCIL </name>
    <id>20006340200016</id>
  </Buyer>
  <CPV>31527200-8</CPV>
  <source>data.gouv.fr_pes</source>
  <duration>5</duration>
  <Pricing>Firm and discountable</Pricing>
  <Price>62210</Price>
  <NotificationDate>2019-07-24+02:00</NotificationDate>
  <id>2019B040100</id>
  <procedure>Adapted procedure</procedure>
  <Suppliers>
    <Supplier>
      <IdentifierType>SIRET</typeIdentifierType>
      <Name>HTB SERVICES</Name>
      <id>83414275400011</id>
    </Supplier>
  </Suppliers>
  <Amendments/>
</PublicContract>
```


b-Example of a dataset collected using the French data portal API

Fig. 5. Example of the collected data.

As the collected data was already structured, no further data extraction was required. Our data collection approach allowed us to retrieve data efficiently and accurately while reducing the resources needed for manual data collection. This enabled us to collect a significant amount of data related to public contracts from 2019 to 2022.

After collecting the data, we utilized Karma, an open-source tool, for semantic transformation. Karma[4] is a semi-automatic tool that implements the approach proposed by [22] for dataset transformation, enabling the definition of mappings from datasets to ontologies, building of semantic models, and publishing of data as RDF. With Karma's machine learning capability, it can learn to map datasets to an ontology by using the attribute values of mapped dataset attributes. When users define relationships between classes, Karma learns from them and can suggest properties and classes to model new sources automatically. Karma's validation feature allows users to verify the proposed model's quality, and it also supports the generation of RML mapping, which automates the RDF graph-based representation of the datasets. Incorporating the PCO ontology in the semantic transformation process enriched the RDF graph output for each collected dataset, representing relationships between data elements, entities, and concepts for constructing the knowledge graph. As a first step in the integration process, all datasets were aligned to the same schema, which was the PCO ontology.

The resulting RDF graphs were combined to form the knowledge graph. We further employed entity linking as a means of completing the integration process. We used the OWL: has key and OWL: Sameas axioms to identify unique discriminative properties for entities in the domain ontology and to link co-referent entities, respectively. For example, we used the SIRET identifier to link unique enterprises as suppliers and the SIREN identifier to link public organizations as public buyers.

Through the use of entity linking, we were able to more effectively integrate data from multiple sources into the knowledge graph, providing a unified and integrated view of data.

The data was stored in an RDF triple store using the Apache Jena framework[5] and made available through Apache Jena Fuseki. This allowed users to query the knowledge graph using the SPARQL query language and retrieve relevant information from the integrated datasets. In total, the knowledge graph consolidated data related to 96,845 public contracts.

E. Knowledge Graph Consumption

To demonstrate the value of our approach, we conducted an analysis using the isolation forest algorithm on integrated public procurement data in our knowledge graph. The isolation forest algorithm is a powerful machine learning tool well-suited to detecting anomalies in datasets. It works by creating a random forest of decision trees and isolating data points that are not consistent with the majority of the data. In the case of public procurement data, the isolation forest algorithm can be used to identify anomalies that may indicate corruption or fraud.

In our study, we used a SPARQL query to extract data from the knowledge graph and applied the isolation forest algorithm to our integrated public procurement data. We found a total of 56 anomalies in the dataset, which were categorized as either favoritism or overpricing. The algorithm identified 27 cases of favoritism and 29 cases of overpricing.

We qualitatively assessed the algorithm results and found that the contracts identified as anomalies were consistently awarded to the same suppliers, indicating a potential case of favoritism. Additionally, we observed that the contract prices for these anomalies were significantly higher than the market price or previous prices for similar contracts, indicating a potential case of overpricing.

The results of this case study demonstrate the value of the proposed approach, which proposes to consolidate data from multiple sources into a single knowledge graph and represent it in a format that promotes analysis and processing for value extraction. By utilizing a knowledge graph and the isolation forest algorithm, we were able to identify potential anomalies that may have gone undetected using a traditional approach.

[4] https://usc-isi-i2.github.io/karma/

[5] https://jena.apache.org/

This can help auditors better understand procurement data, detect potential fraud or corruption, and ultimately promote a more transparent and accountable government.

V. CONCLUSION

In this paper, we introduce an approach for extracting value from open government data and transforming it into valuable knowledge. Our approach utilizes a layered architecture to extract and collect structured and unstructured data sources, transform them into a machine-processable, understandable, and semantically rich format, and integrate them into a knowledge graph. The proposed approach offers an efficient and generic way to construct the graph, which can be further utilized for a plethora of applications. The approach was implemented in the context of public procurement and successfully identified anomalies such as overpricing and favoritism. Although the paper is focused on public procurement, the proposed approach applies to other governmental domains. As a future direction, we aim to propose a detailed description of how to exploit the knowledge graph, to further improve the effectiveness of the proposed approach.

REFERENCES

[1] Attard, J., Orlandi, F., Scerri, S., & Auer, S. (2015). A systematic review of open government data initiatives. Government information quarterly, 32(4), 399-418.

[2] Ubaldi, B. (2013). Open government data: Towards empirical analysis of open government data initiatives.

[3] Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. Information systems management, 29(4), 258-268

[4] K. Y. Dahbi, H. Lamharhar, D. Chiadmi Exploring dimensions influencing the usage of Open Government Data portals. In Proceedings of the 12th International Conference on Intelligent Systems: Theories and Applications (pp. 1-6).

[5] S. Neumaier, J. Umbrich, A. Polleres (2016). Automated quality assessment of metadata across open data portals. Journal of Data and Information Quality (JDIQ), 8(1), 1-29.

[6] K. Y. Dahbi, H. Lamharhar, D. Chiadmi .Toward a user-centered approach to enhance Data discoverability on Open Government Data portals. In 2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS) (pp. 1-5). IEEE.

[7] M. Mountantonakis, Y. Tzitzikas, (2019). Large-scale semantic integration of linked data: A survey. ACM Computing Surveys (CSUR), 52(5), 1-40. 8. H.Purohit , R. Kanagasabai, N. Deshpande (2019, January).

[8] Towards Next Generation Knowledge Graphs for Disaster Management. In 2019 IEEE 13th International Conference on Semantic Computing (ICSC) (pp. 474-477). IEEE.

[9] Y. Jia, Y.Qi , H.Shang , R. Jiang, A. Li, (2018). A practical approach to constructing a knowledge graph for cybersecurity. Engineering, 4(1), 53-60. 10. M. Wang, Q. Zeng, W. Chen , J. Pan, H. Wu, C. Sudlow, D. Robertson, (2020).

[10] Building the Knowledge Graph for UK Health Data Science.

[11] J. M. Gomez-Perez, J. Z .Pan, G. Vetere, , H. Wu, (2017). Enterprise knowledge graph: An introduction. In Exploiting linked data and knowledge graphs in large organisations (pp. 1-14). Springer, Cham.

[12] Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., Melo, G. D., Gutierrez, C., ... & Zimmermann, A. (2021). Knowledge graphs. ACM Computing Surveys (CSUR), 54(4), 1-37.

[13] A. Hogan, E . Blomqvist, M .Cochez, C . d'Amato, G. de Melo, C. Gutierrez, , R. Navigli(2020). Knowledge graphs. arXiv preprint arXiv:2003.02320. 14. J. Yan, C. Wang, , W. Cheng, M. Gao, A. Zhou, (2018). A retrospective of knowledge graphs. Frontiers of Computer Science, 12(1), 55-74. https://doi.org/10.1007/s11704-016-5228-9

[14] J. Yan, C. Wang, , W. Cheng, M. Gao, A. Zhou, (2018). A retrospective of knowledge graphs. Frontiers of Computer Science, 12(1), 55-74. https://doi.org/10.1007/s11704-016-5228-9

[15] S. Ronzhin, E. Folmer, P. Maria, M. Brattinga, W. Beek, R. Lemmens, R., R. van't Veer(2019). Kadaster knowledge graph: Beyond the fifth star of open data. Information, 10(10), 310.

[16] P. Espinoza-Arias, M. J. Fernández-Ruiz, V.Morlán-Plo, R.Notivol-Bezares, O. Corcho (2020). The Zaragoza's Knowledge Graph: Open Data to Harness the City Knowledge. Information, 11(3), 129.

[17] Rafail, P., & Efthimios, T. (2020, November). Knowledge Graphs for Public Service Description: The Case of Getting a Passport in Greece. In European, Mediterranean, and Middle Eastern Conference on Information Systems (pp. 270-286). Springer, Cham.

[18] Soylu, A., Corcho, O., Elvesæter, B., Badenes-Olmedo, C., Blount, T., Yedro Martínez, F., ... & Roman, D. (2022). TheyBuyForYou platform and knowledge graph: Expanding horizons in public procurement with open linked data. Semantic Web, (Preprint), 1-27.

[19] Soylu, A., Elvesæter, B., Turk, P., Roman, D., Corcho, O., Simperl, E., ... & Lech, T. C. (2019). An overview of the TBFY knowledge graph for public procurement. CEUR Workshop Proceedings

[20] Bratsas, C., Chondrokostas, E., Koupidis, K., & Antoniou, I. (2021). The use of national strategic reference framework data in knowledge graphs and data mining to identify red flags. Data, 6(1), 2.

[21] Cifuentes-Silva, F., Fernández-Álvarez, D., & Labra-Gayo, J. E. (2020). National budget as linked open data: New tools for supporting the sustainability of public finances. Sustainability, 12(11), 4551.

[22] Taheriyan, M., Knoblock, C. A., Szekely, P., & Ambite, J. L. (2016). Learning the semantics of structured data sources. Journal of Web Semantics, 37, 152-169.

[23] Dimou, A., Vander Sande, M., Colpaert, P., Verborgh, R., Mannens, E., & Van de Walle, R. (2014, January). RML: a generic language for integrated RDF mappings of heterogeneous data. In Ldow.

[24] Oliveira, I. L., Fileto, R., Speck, R., Garcia, L. P., Moussallem, D., & Lehmann, J. (2021). Towards holistic entity linking: Survey and directions. Information Systems, 95, 101624.

[25] Papadakis, G., Skoutas, D., Thanos, E., & Palpanas, T. (2019). A survey of blocking and filtering techniques for entity resolution. arXiv preprint arXiv:1905.06167.

[26] Kundu, O., James, A. D., & Rigby, J. (2020). Public procurement and innovation: a systematic literature review. Science and Public Policy, 47(4), 490-502.

[27] Rustiarini, N. W., Nurkholis, N., & Andayani, W. (2019). Why people commit public procurement fraud? The fraud diamond view. Journal of public procurement, 19(4), 345-362.

[28] Baranek, B., & Titl, V. (2020). The cost of favoritism in public procurement. FEB Research Report Department of Economics.

[29] Modrušan, N., Rabuzin, K., & Mršic, L. (2021). Review of Public Procurement Fraud Detection Techniques Powered by Emerging Technologies. International Journal of Advanced Computer Science and Applications, 12(2).

[30] Nečaský, M., Klímek, J., Mynarz, J., Knap, T., Svátek, V., & Stárka, J. (2014). Linked data support for filing public contracts. Computers in Industry, 65(5), 862-877.