Enhancing Administrative Source Registers for the Development of a Robust Large Language Model: A Novel Methodological Approach

Adham Kahlawi, Cristina Martelli
Department of Statistics, Computer Science, Applications, University of Florence, Florence, Italy

***Abstract*—Accurate statistical information is critical for understanding, describing, and managing socio-economic systems. While data availability has increased, often it does not meet the quality requirements for effective governance. Administrative registers are crucial for statistical information production, but their potential is hampered by quality issues stemming from administrative inconsistencies. This paper explores the integration of semantic technologies, including ontologies and knowledge graphs, with administrative databases to improve data quality. We discuss the development of large language models (LLMs) that enable a robust, queryable framework, facilitating the integration of disparate data sources. This approach ensures high-quality administrative data, essential for statistical reuse and the development of comprehensive, dynamic knowledge graphs and LLMs tailored for administrative applications.**

Keywords—Statistical information systems; administrative data reuse; ontology; database; semantic web; knowledge graph; LLM

I. Introduction

In the face of the increasing complexity of modern societies, innovative governance and decision-making approaches are essential to navigate the evolving socio-economic and political landscapes [1]. Recent shifts towards the autonomy of local actors, the creation of new institutional arenas, global economic repositioning, decentralization, and a transition to network societies have underscored the importance of network structures over hierarchical ones, creating a demand for knowledge bases capable of understanding and managing these complexities [2]. Administrative data, defined comprehensively by Eurostat as data collected for non-statistical programs by both governmental and private organizations, have emerged as a focal point. The strategic use of this data can generate administrative and statistical information, which serves as both a tool for harmonizing administrative processes and decision-making, and a means of communication within and outside organizations [3].

The essence of governing complex socio-economic environments lies in the profound understanding of their actors, relationships, and processes, necessitating systems that are deeply rooted in reality and supported by active observation [4]. The Organization for Economic Cooperation and Development highlights the role of administrative data as a reliable source for statistical information, emphasizing the importance of their collection, processing, and storage. Technological advancements have enhanced data production processes, presenting new opportunities and challenges in data utilization, transparency, and integration [5].

Typically, the details about the administrative sources are dispersed across various isolated databases created by different departments and divisions. This fragmentation prevents Public Administrations from offering a thorough understanding of their key entities and their interactions. Lately, knowledge graphs have emerged as a solution to organize vast data sets effectively, but they primarily depict a fixed snapshot of the world. They often overlook the dynamic aspects and the evolution over time [6].

This paper suggests an approach grounded in semantic web technology to develop administrative systems, designed to be statistically reusable and ready to be represented as a graph structure. This aims to create administrative sources suitable for querying Linked Data Models (LLMs), capable of bridging the gap between administrative and statistical information and ready to be integrated with various sources. This effort addresses the challenges of generating big data and reusing statistical data.

This approach recognizes the inherent limitations of current big data management practices, such as issues of data quality, coverage, and cost, and seeks to overcome these by leveraging administrative data as a mean to describe the granularity, complexity and interconnectivity of reality. By focusing on the early stages of the data production process and employing new technologies for data integration and modeling [7, 8], this study aims to ensure that administrative data are not only valuable in their own right but also fit for statistical reuse and capable of representing the socio-economic complexity of our world.

Finally, we can summarize the core of this paper in four points:

- Problem Statement and Questions:

Administrative data, collected by both governmental and private organizations, have emerged as crucial but are often marred by inconsistencies and low quality, impeding their full utilization for governance and decision-making. How can we improve the quality and integration of administrative data to better support complex governance needs?

- Objectives:

This study aims to explore the potential of semantic technologies in enhancing the quality and utility of administrative data. By integrating these technologies with administrative databases, we seek to develop a robust method for producing high-quality administrative data that is statistically reusable and supports complex decision-making processes.

- Significance:

The strategic use of improved administrative data could revolutionize decision-making processes, providing a more coherent and dynamic understanding of socio-economic environments. This could lead to more informed policies and efficient governance systems.

- Presentation of the Study:

In this paper, we propose a novel methodological approach using semantic web technologies to address the challenges associated with administrative data. Our approach not only enhances data quality but also facilitates the integration of diverse data sources, laying a foundational structure for creating LLM capable of bridging the gap between administrative records and statistical information needs.

The structure of this paper is as follows: Section II presents a review of the literature relevant to our study; Section III outlines the methodology we adopted; Section IV discusses the application of this methodology and the results obtained; and Section V concludes the paper with a summary of our findings and suggestions for future research.

II. Related Works

In the realm of enhancing data quality and interlinking public datasets with knowledge graphs (KGs), notable contributions have emerged, offering innovative approaches and methodologies. Among these, the work presented in study [9] by Haklae Kim stands out by addressing the challenges of utilizing government codes within public datasets. The paper highlights how government codes, crucial for standardizing administrative procedures, often become obscured when included in public data, thereby limiting their utility and impeding dataset interlinking. Kim proposes employing the administrative codes generated by the Korean government as a standard in public data environments, leveraging an ontology model to encapsulate the data structure and meaning of administrative codes. This approach, through the construction of a comprehensive knowledge graph, seeks to enhance the quality and connectivity of coded information in public datasets, thus facilitating standardized access to administrative codes beyond government systems.

Similarly, [10] by Dimitris Zeginis and Konstantinos Tarabanis introduces an event-centric knowledge graph (ECKG) model to improve data governance and analysis within public administrations (PAs). Recognizing the vast amounts of data generated by PAs and their often fragmented nature across different databases, Zeginis and Tarabanis pinpoint a gap in existing KG models that tend to represent static data, neglecting the dynamic nature of data interactions. By prioritizing events as primary entities for knowledge representation, their model aims to capture the dynamic aspects of public service interactions, offering a more comprehensive overview of interactions between core entities like citizens, businesses, and PAs themselves. This method not only facilitates citizen-friendly public administration but also enables advanced data analytics and AI applications by integrating data both in representation and in context. Expanding on this concept, [6] by the same authors applies the ECKG model to the Greek PA, demonstrating its potential to provide a comprehensive view of public administrations (PA) interactions, support data analytics, and aid in real-time decision-making. This model uses Core Public Service Vocabulary Application Profile (CPSV-AP) to describe public services, distinguishing between event-aware and event-agnostic concepts, thereby efficiently managing public service versions and variants. A case study on the " birth registration " life event in Greece showcases how the ECKG model can capture PA interactions' complexity, enhancing data integration, analytics, and providing a 360-degree view of end-users.

In the healthcare domain, [11] by Arif Khan, Shahadat Uddin, and Uma Srinivasan utilizes administrative health data to predict the risk of Type 2 Diabetes (T2D). Applying data mining and network analysis techniques on a dataset comprising 1.4 million records from 0.75 million patients, the study develops a prediction framework that enhances prediction accuracy through innovative graph theory and social network-based measures. This approach offers a cost-effective method for healthcare providers and insurers to identify high-risk cohorts for preventive strategies, aiming to mitigate the burden of chronic diseases on healthcare resources.

In the data integration domain, [12] by Enayat Rajabi, Rishi Midha, and Jairo Francisco de Souza address the challenge of integrating disparate datasets within open government data portals. Through the use of Semantic Web technologies and the transformation of datasets into the Resource Description Framework (RDF) format, the authors illustrate the benefits of applying Semantic Web standards to government datasets, enabling sophisticated querying capabilities. Also, in study [13] by Luis M. Vilches-Blázquez and Jhonny Saavedra introduce a pioneering approach for the integration and management of heterogeneous land administration data through graph-based knowledge representation. The study acknowledges the considerable challenges arising from the variety of data formats, models, and standards spread across different Colombian land administration agencies. To address these challenges, the authors propose an ontology-based framework that aligns with both national and international standards for land administration. This innovative framework is designed to promote the harmonization, interoperability, sharing, and integration of data across decentralized and multi-jurisdictional agencies without necessitating modifications to their existing processes, models, or vocabularies. Employing a methodology that constructs knowledge graphs based on ontology, the framework connects various datasets through a unified identifier for land administration features and enriches these graphs with spatial connections and data sourced from the Linked Open Data cloud. Through a case study focusing on the integration of data from the Colombian National Geographic Institute (IGAC) and the Bogota cadastre, the paper effectively demonstrates how knowledge graphs can address semantic heterogeneity and enhance the management and utilization of data in land administration.

Some studies have explored the relationship between Knowledge Graph and Large Language Models such as: [14] by Qing Huang et al. explores enhancing API recommendation systems by integrating Large Language Models (LLMs) guided by a Knowledge Graph (KG). This study tackles the challenges of utilizing government codes within public datasets by proposing the use of administrative codes as a standard in public data environments. By employing an ontology model to represent the data structure and meaning of these codes, the research assesses the accuracy and connectivity of administrative codes in public data, showing potential for enhancing the quality and connectivity of coded information in public datasets. Also, [15] by Shuang Yu, Tao Huang, Mingyi Liu, and Zhongjie Wang introduces BEAR, a service domain KG constructed to address the lack of large-scale, high-quality KGs in the service computing community. Utilizing LLMs for zero-shot knowledge extraction and guided by a well-designed service domain ontology, BEAR demonstrates significant advancements in domain-specific KG construction methodologies, containing over 130,000 entities, 160,000 relations, and approximately 424,000 factual knowledge attributes. This construction process leverages the semantic understanding and reasoning capabilities of LLMs to overcome challenges related to data scarcity and complexity, highlighting the potential to drive application and algorithm innovation within the service computing field. Additionally, [16] by Linyao Yang et al. explores the enhancement of large language models (LLMs) with knowledge graphs (KGs) to improve factual accuracy in text generation. Categorizing methods into before-training, during-training, and post-training enhancements, the paper advocates for the combination of KGs and LLMs to address factual reasoning limitations, suggesting new research avenues. Furthermore, [17] by Shirui Pan et al. offers a comprehensive framework for the integration of large language models (LLMs) like GPT-4 with knowledge graphs (KGs), aiming to augment the capabilities of both technologies and mitigate their individual limitations. The roadmap presented in the paper is organized around three core frameworks: KG-enhanced LLMs, LLM-augmented KGs, and a synergized integration of LLMs and KGs. The first framework, KG-enhanced LLMs, is focused on embedding KGs into the training and inference phases of LLMs to supply external knowledge, thereby improving inference and enhancing interpretability. The second framework, LLM-augmented KGs, leverages the computational power of LLMs to address challenges in KG tasks, including embedding, completion, construction, and question answering, which are often hindered by incompleteness and the difficulty of incorporating new knowledge. Lastly, the synergized framework proposes a bidirectional enhancement strategy, whereby LLMs and KGs mutually benefit from each other, thus fostering advanced knowledge representation and reasoning capabilities. This roadmap meticulously categorizes research efforts within these frameworks, explores emerging advancements, and outlines the challenges and future directions, underscoring the significant potential of merging LLMs' proficiency in language processing with the structured knowledge representation of KGs for a variety of applications. Lastly, [18] by Amir Hassan Shariatmadari et al., and "Unifying Large Language Models and Knowledge Graphs: A Roadmap" by Shirui Pan et al. both emphasize the synergy between LLMs and KGs. The former investigates the use of Cross-Modal Attention mechanisms to improve LLM explainability in the biomedical domain, while the latter presents a roadmap for integrating LLMs and KGs to enhance their collective capabilities, identifying challenges and future directions in knowledge representation and reasoning across various applications. These studies collectively highlight the evolving landscape of knowledge representation, emphasizing the significant potential of integrating diverse methodologies to address complex challenges in data analysis, management, and utilization.

III. Methodology

Fig. 1 depicts the methodology, structured into four steps (represented by the horizontal segments with arrows and identified by their respective numbers). Each step unfolds through one or more activities.

A. *Domain Analysis*

The figure depicts a scenario where administrative data is stored in multiple administrative databases. The first activity (identified by segment -1-) will therefore involve extracting the structure of the databases by analyzing the tables, their relationships, along with the names of all the columns within the tables.

In the second activity (segment -2-) we examine all the columns in the databases and sort them into three groups. The first group consists of columns that can be assigned or derived from a classification. The second group contains columns in which the information can be standardized into fixed options. Lastly, the third group comprises all columns that do not fall into either of the previous two groups.

B. *Create the Domain Ontology*

The creation of the ontology for the domain served by the considered administrative sources begins by representing in terms of ontologies the concepts represented in the columns of the analyzed databases.

In the third activity (segment -3-), we begin by creating an ontology for each classification that represents one of the columns within the standard classification group. Next, we create an ontology for each standard concept which is identified through the columns in the standard concept group.

After having translated the columns of the databases into ontological terms, the next step involves constructing the overall ontology of the domain.

Moving on to the fourth activity (segment -4-) we start by building an ontology of the administrative resource from the database schema. This is done by following four rules:

1) Create a class for each table in the database, unless the table contains only foreign keys.

2) Create an object property to connect two classes that are related through their tables.

3) Create an object property for each column belonging to the first and second group, to connect its table's class with the ontology created from it.

4) Create a data property for each table's column that belongs to the third group.

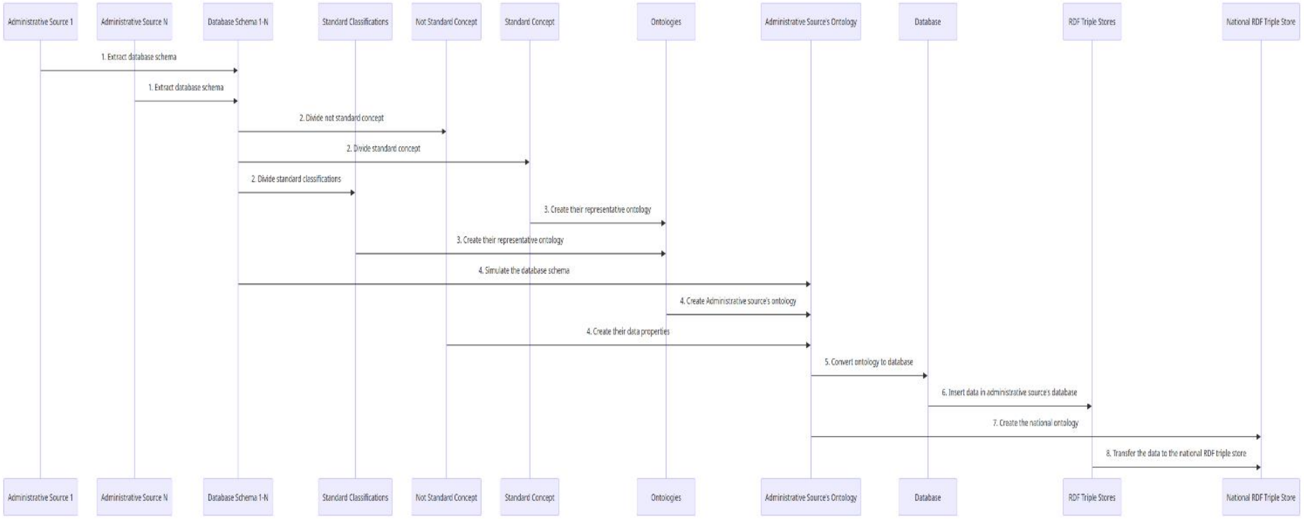

Administrative Source 1
Administrative Source N
Database Schema 1-N
Standard Classifications
Not Standard Concept
Standard Concept
Ontologies
Administrative Source's Ontology
Database
RDF Triple Stores
National RDF Triple Store
1. Extract database schema
1. Extract database schema
2. Divide not standard concept
2. Divide standard concept
2. Divide standard classifications
3. Create their representative ontology
3. Create their representative ontology
4. Simulate the database schema
4. Create Administrative source's ontology
4. Create their data properties
5. Convert ontology to database
6. Insert data in administrative source's database
7. Create the national ontology
8. Transfer the data to the national RDF triple store
Administrative Source 1
Administrative Source N
Database Schema 1-N
Standard Classifications
Not Standard Concept
Standard Concept
Ontologies
Administrative Source's Ontology
Database
RDF Triple Stores
National RDF Triple Store

Fig. 1. Methodology of proposed model.

C. Generate the New Database

In the fifth activity (segment -5-), we will generate a new database by utilizing the ontology created in the fourth activity. In this process, each class in the ontology will be transformed into a table in the database. Similarly, each data property will be turned into a column in the table, and each object property will establish a relationship between the tables.

However, it's important to keep in mind that the conversion process from an ontology to a database requires several controls to ensure its accuracy and efficacy. The following rules apply while converting an ontology into a table:

1) The classes that have an available number of individuals will be converted to a table; in contrast, if these classes were part of a hierarchical classification, their individuals would be fused together and placed in a table that takes the higher class's name in this classification;

2) All the classes with a limited number of individuals will be converted to a column in the table that was the domain class for the object property that was the range class for it;

3) All the data properties will be converted to a column in the table that was the domain class for it; and

4) The object properties can be converted in three ways depending on the restriction type used in the ontology. In the first type, restrictions are placed between the two classes on a one-to-one basis. In the second type, restrictions are placed on a one-to-many basis. In the third way, the two classes are related on a many-to-many basis; in this case, a new table will be created with two columns, one with the primary key of the domain class table and the second with the primary key of the range class table. The relationship between these three methods can be described as domain-to-now table, one-to-many table same as for now able-to-range table.

The new database and ontology will be fully compatible with each other to operate together, and this compatibility will be ensured once the new system is in use.

The sixth activity of the system allows the data received from the users to be inserted into the database and the triple store simultaneously, without needing any input from the user.

This is an important step to document, through the semantic web, any innovative administrative archives best practice, while also dynamically updating the domain ontology in response to field developments.

The first way involves inserting the data into the database, which will be used to manage the administrative source and meet its needs. The second way involves inserting the data into a triple store, which is the physical location where the data is stored. The data stored in the triple store is machine-readable and machine-understandable, as it is written in the Resource Description Framework language. This language represents any resource with a unique Uniform Resource Identifier (URI), even if that resource is found in different domains. The triple store has a simple structure, like a text file, which means that adding new data to it does not require any pre-processing and can be combined with previous data by the computer without human intervention, using URIs. This triple store will be used in the next activity to integrate data from various administrative sources.

D. Generate the National Ontology and the National Triple Store

In the process of creating a national database from different administrative sources, the seventh activity involves merging the ontologies of these sources. This step is made possible by the presence of shared classes that represent standard classifications and concepts developed in the second step. The merging process is automated. The eighth activity involves building a national triple store by collecting data from different administrative sources. This data will be used for statistical studies to support decision-makers at local and national levels.

IV. Methodology Application and Result

During the construction of the Italian high-speed trains, a methodology was developed to integrate different and diverse data sources, derived from partial and not harmonized administrative views of the problematic area. The purpose of this study was twofold. Firstly, it aimed to set up a methodology to integrate different administrative information systems that were already on the site, yet were heterogeneous and not integrable. Secondly, it aimed to build an asset for automatically generating an administrative database that is statistically reusable by design.

The first objective is derived from the experience of constructing the knowledge base for the Italian high-speed trains construction sites. The administrative information systems were already on site, but they were not integrable. Therefore, the methodology to integrate these data sources had to be developed. The second objective is to propose a similar information system to different construction sites, so that management processes can be carried out effectively and the produced data can be immediately reusable without the efforts paid in the first experience.

Both of these objectives rely on a construction site ontology, which is the starting point for generating the administrative database. In the first case, the construction site ontology is derived from the already existing and not integrable databases. In the second case, the ontology is used to generate the administrative database.

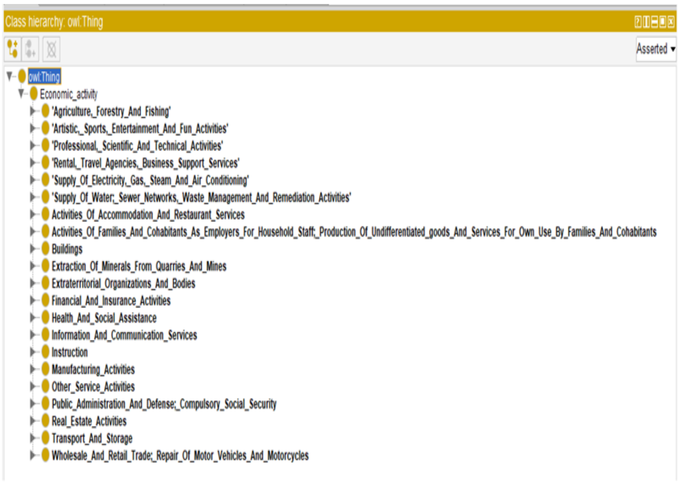
In this study, we will clarify the process of applying the methodology to one of the database tables, and in the appendix 1 we will explain the results of the conversion process for the entire database.

We analyzed Table I and found columns for standard classifications and concepts.

Consequently, as we mention in the second step of the methodology, we will create sub-ontologies for all standard classification and concept categories. Fig. 2 illustrates the ontology of the Nomenclature of Economic Activities NACE.

TABLE I. The Analysis of the Firm's Table

Firm's table	
columns name	analysis result
VAT number	Primary Key
Name	the general concept category
Economic Activity Codes	the standard classification category "Atico2007"
INAIL Rate Codes	the standard classification category "Inail"
street of Registered Office	the general concept category
Postal Code of Registered Office	the standard classification category
City of Registered Office	the standard classification category
province of Registered Office	the standard classification category
Region of Registered Office	the standard classification category



```
owl:Thing
  Economic_activity
    'Agriculture,_Forestry_And_Fishing'
    'Artistic,_Sports,_Entertainment_And_Fun_Activities'
    'Professional,_Scientific_And_Technical_Activities'
    'Rental,_Travel_Agencies,_Business_Support_Services'
    'Supply_Of_Electricity,_Gas,_Steam_And_Air_Conditioning'
    'Supply_Of_Water;_Sewer_Networks,_Waste_Management_And_Remediation_Activities'
    Activities_Of_Accommodation_And_Restaurant_Services
    Activities_Of_Families_And_Cohabitants_As_Employers_For_Household_Staff;_Production_Of_Undifferentiated_goods_And_Services_For_Own_Use_By_Families_And_Cohabitants
    Buildings
    Extraction_Of_Minerals_From_Quarries_And_Mines
    Extraterritorial_Organizations_And_Bodies
    Financial_And_Insurance_Activities
    Health_And_Social_Assistance
    Information_And_Communication_Services
    Instruction
    Manufacturing_Activities
    Other_Service_Activities
    Public_Administration_And_Defense;_Compulsory_Social_Security
    Real_Estate_Activities
    Transport_And_Storage
    Wholesale_And_Retail_Trade;_Repair_Of_Motor_Vehicles_And_Motorcycles
```

Fig. 2. Ontology of economic activities NACE.

In the third step of creating the case study, we develop an ontology based on the database structure and the ontologies created in the previous step. Fig. 3 illustrates how the firm's table was transformed into four primary classes. Each of these classes contains subsets. For instance, the Italian Address class comprises four subclasses representing the structure of any address, including street name and building number. During this transformation, we considered data property, and economic activities contain hierarchical subclasses defined by 109 subclasses.

The individuals of the four classes will have different object properties governing their relationships. However, the relationship between the "firm" and "Italian Address" will be governed by the individuals of the subclass "postal code". This is because the relationship between the individuals of the "Italian Address" subclasses is fixed and was defined at the time of the creation of the ontology.

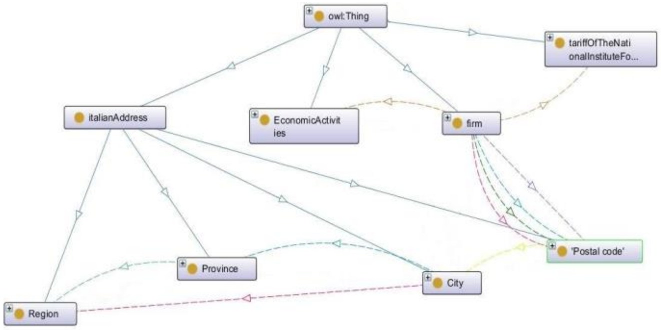


Fig. 3. Ontological representation of firms' table.

The Appendix 2 displays the ontological structure that depicts the ontology of the case study.

This ontology defines and specifies every concept used in the original database, supports creation of a harmonised language among different construction sites, and can be translated into a coherent relational database.

We will proceed to create a new database from the ontology constructed in the second step, taking into account all the rules established in the methodology. The Appendix 3 presents the structure of the database obtained directly from the ontology. Comparing it with the original database, we can see that it has a similar structure, but it is more efficient since it has undergone a rigorous process of creating classes and properties. As a result, proactive service is provided to the institution whose data is reused in an information system to support decision-making. Instead of burdening them with coding, we provide them with the structure to be used directly.

When the new database is used to manage the domain, it will effectively organize and manage its data, as well as generate a triple store. After that, the integrated ontology and the integrated triple store will be created automatically, taking advantage of what was built in the previous steps. The integrated ontology is an ontology that is created by integrating a group of ontologies that represent different domains. The connection between these ontologies is made through the sub-ontologies established in the first step. The integrated triple store is the triple store that is created by integrating a group of triple stores, which are developed by applying the previous steps to different domains.

All in all, the prior related works provided only partial solutions and did not address issues at a holistic level, unlike the methodology implemented in this research, which aims to offer a comprehensive solution. Additionally, this work focuses on improving the existing data collection system to enhance the management of administrative resources. It also involves developing a parallel system that ensures the integration of data from various administrative sources in a cohesive manner and guarantees the effective reusability of the data.

V. Conclusion

The emergence of new semantic technologies presents a challenge, an opportunity, and a risk to official statistics. On one hand, these technologies offer unprecedented processing power to manage quantitative information; on the other hand, there is a risk of generating information systems that fall short of the quality standards necessary for statistical analysis.

In this study, we explore the statistical reuse of administrative sources in light of the potential for conscious integration with semantic technology. By rethinking the reuse of administrative data, we can contain the key waste of public memory that arises from the difficulty of integrating sources. We need information systems that are suitable for managing problems and services, but also support the reuse of their data.

While big data methodologies exist and are increasingly popular, they may not provide the necessary level of detail, quality, and precision required for specific and delicate domains, such as the ones served by PA. Therefore, we focus on using semantic web technologies to support the entire process of generating archives, starting from the moment of their conceptualization.

Our study shows that semantic web technologies can be used to accurately analyze and describe a domain, which can help build a high-quality database. They can also aid in integrating data from different sources without the need for manual intervention, allowing for the reuse of data for statistical and non-statistical purposes simultaneously. The national RDF triple store represents the national administrative knowledge graph that we can use to create or fine-tune the administrative LLM.

Our study also highlights the unprecedented areas of presence for statistical agencies, such as the supervision of language and conceptualizations. Adopting these methods on a broader scale would lead to a different quality of administrative sources. This integration not only supports the broader dissemination of official codifications but also recognizes the methods of experts from different domains, allowing for their integration and official dissemination.

The possibility of connoting each concept with an official identifier stored on the internet, the choice of having these methods adopted by social and economic actors, and the constitution of large texts that can be interpreted automatically shifts the usual horizons of those who deal with statistical information systems. This creates new challenges for the statistical community, such as processes for linkage or testing the conditions of respect for privacy.

In future work, we will seek to create an administrative resource LLM and establish the mechanism for its use and the controls that will govern this use.

References

[1] N. Stehr, Modern Societies as Knowledge Societies BT - Nico Stehr: Pioneer in the Theory of Society and Knowledge, in: M.T. Adolf (Ed.), Springer International Publishing, Cham, 2018: pp. 309–331. https://doi.org/10.1007/978-3-319-76995-0_20.

[2] C.B. Keating, P.F. Katina, Complex system governance: Concept, utility, and challenges, Syst. Res. Behav. Sci. 36 (2019) 687–705. https://doi.org/https://doi.org/10.1002/sres.2621.

[3] H.M. dos Santos, G.L. Krawszuk, Organizational knowledge management: archival processing for reuse of administrative information, Investig. Bibl. Arch. Bibl. e Inf. Vol 34, No 83 (2020)DO - 10.22201/Iibi.24488321xe.2020.83.58146 . (2020). http://rev-ib.unam.mx/ib/index.php/ib/article/view/58146.

[4] L.A.A. Terra, J.L. Passador, Strategies for the Study of Complex Socio-Economic Systems: an Approach Using Agent-Based Simulation, Syst. Pract. Action Res. 31 (2018) 311–325. https://doi.org/10.1007/s11213-017-9427-6.

[5] OECD, Measuring the Non-Observed Economy: A Handbook, Organisation for Economic Co-operation and Development, 2002. https://doi.org/https://doi.org/https://doi.org/10.1787/9789264175358-en.

[6] D. Zeginis, K. Tarabanis, An Event-Centric Knowledge Graph Approach for Public Administration as an Enabler for Data Analytics, Computers. 13 (2024). https://doi.org/10.3390/computers13010017.

[7] A. Kahlawi, An Ontology-driven DBpedia Quality Enhancement to Support Entity Annotation for Arabic Text, Int. J. Adv. Comput. Sci. Appl. 14 (2023). https://doi.org/10.14569/IJACSA.2023.0140301.

[8] A. Kahlawi, An Ontology Driven ESCO LOD Quality Enhancement, Int. J. Adv. Comput. Sci. Appl. 11 (2020). https://doi.org/10.14569/IJACSA.2020.0110308.

[9] H. Kim, Knowledge Graph of Administrative Codes in Korea: The Case for Improving Data Quality and Interlinking of Public Data, J. Inf. Sci. THEORY Pract. 11 (2023). https://doi.org/https://doi.org/10.1633/JISTaP.2023.11.3.4.

[10] D. Zeginis, K. Tarabanis, Towards an event-centric knowledge graph approach for public administration, in: 2022 IEEE 24th Conf. Bus. Informatics, 2022: pp. 25–32. https://doi.org/10.1109/CBI54897.2022.10045.

[11] A. Khan, S. Uddin, U. Srinivasan, Chronic disease prediction using administrative data and graph theory: The case of type 2 diabetes, Expert Syst. Appl. 136 (2019) 230–241. https://doi.org/https://doi.org/10.1016/j.eswa.2019.05.048.

[12] . Rajabi, R. Midha, J.F. de Souza, Constructing a knowledge graph for open government data: the case of Nova Scotia disease datasets, J. Biomed. Semantics. 14 (2023) 4. https://doi.org/10.1186/s13326-023-00284-w.

[13] L.M. Vilches-Blázquez, J. Saavedra, A graph-based representation of knowledge for managing land administration data from distributed agencies – A case study of Colombia, Geo-Spatial Inf. Sci. 25 (2022) 259–277. https://doi.org/10.1080/10095020.2021.2015250.

[14] Q. Huang, Z. Wan, Z. Xing, C. Wang, J. Chen, X. Xu, Q. Lu, Let's Chat to Find the APIs: Connecting Human, LLM and Knowledge Graph through AI Chain, in: 2023 38th IEEE/ACM Int. Conf. Autom. Softw. Eng., 2023: pp. 471–483. https://doi.org/10.1109/ASE56229.2023.00075.

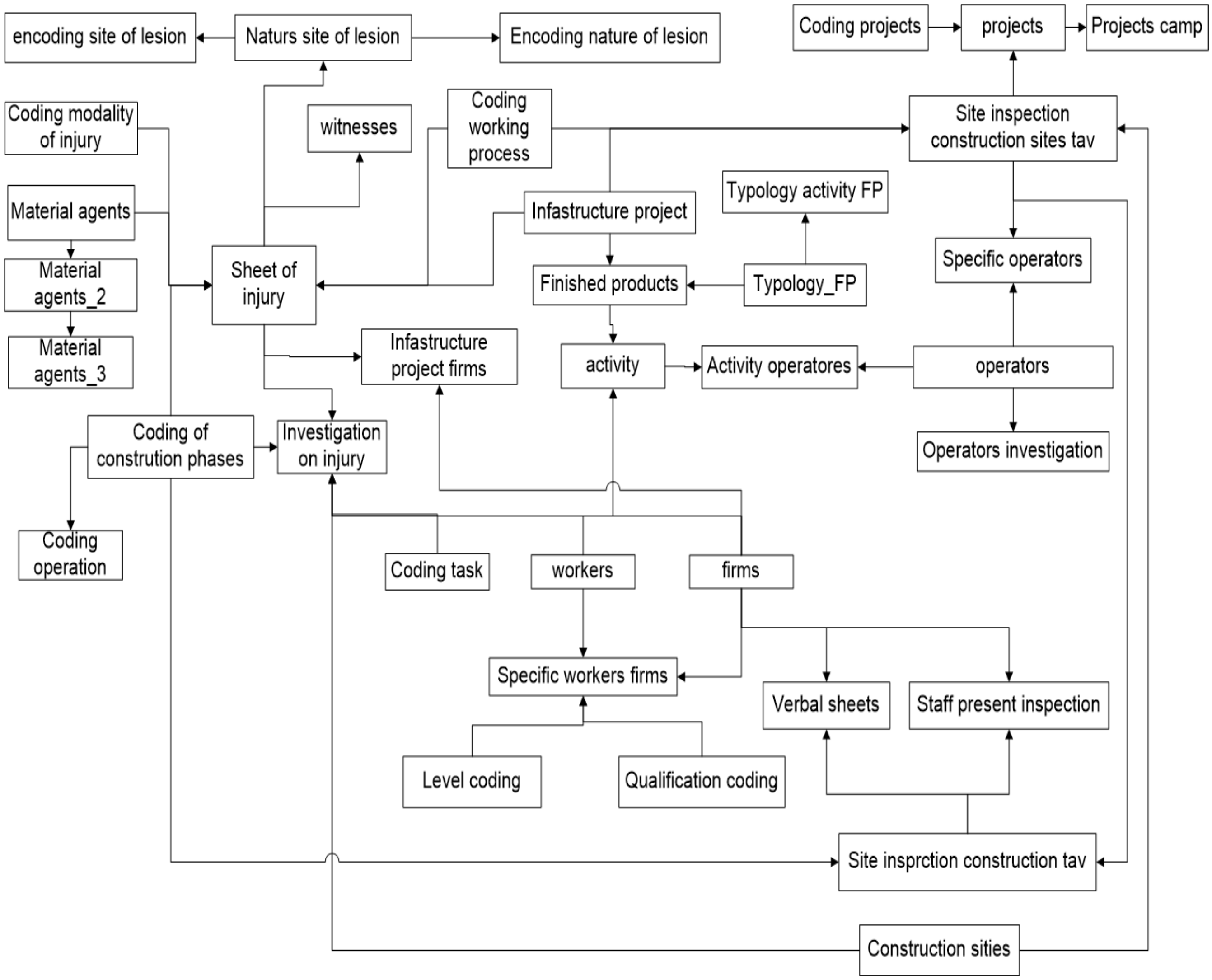
[15] S. Yu, T. Huang, M. Liu, Z. Wang, BEAR: Revolutionizing Service Domain Knowledge Graph Construction with LLM, in: F. Monti, S. Rinderle-Ma, A. Ruiz Cortés, Z. Zheng, M. Mecella (Eds.), Serv. Comput., Springer Nature Switzerland, Cham, 2023: pp. 339–346.

[16] L. Yang, H. Chen, Z. Li, X. Ding, X. Wu, Give Us the Facts: Enhancing Large Language Models with Knowledge Graphs for Fact-aware Language Modeling, IEEE Trans. Knowl. Data Eng. (2024) 1–20. https://doi.org/10.1109/TKDE.2024.3360454.

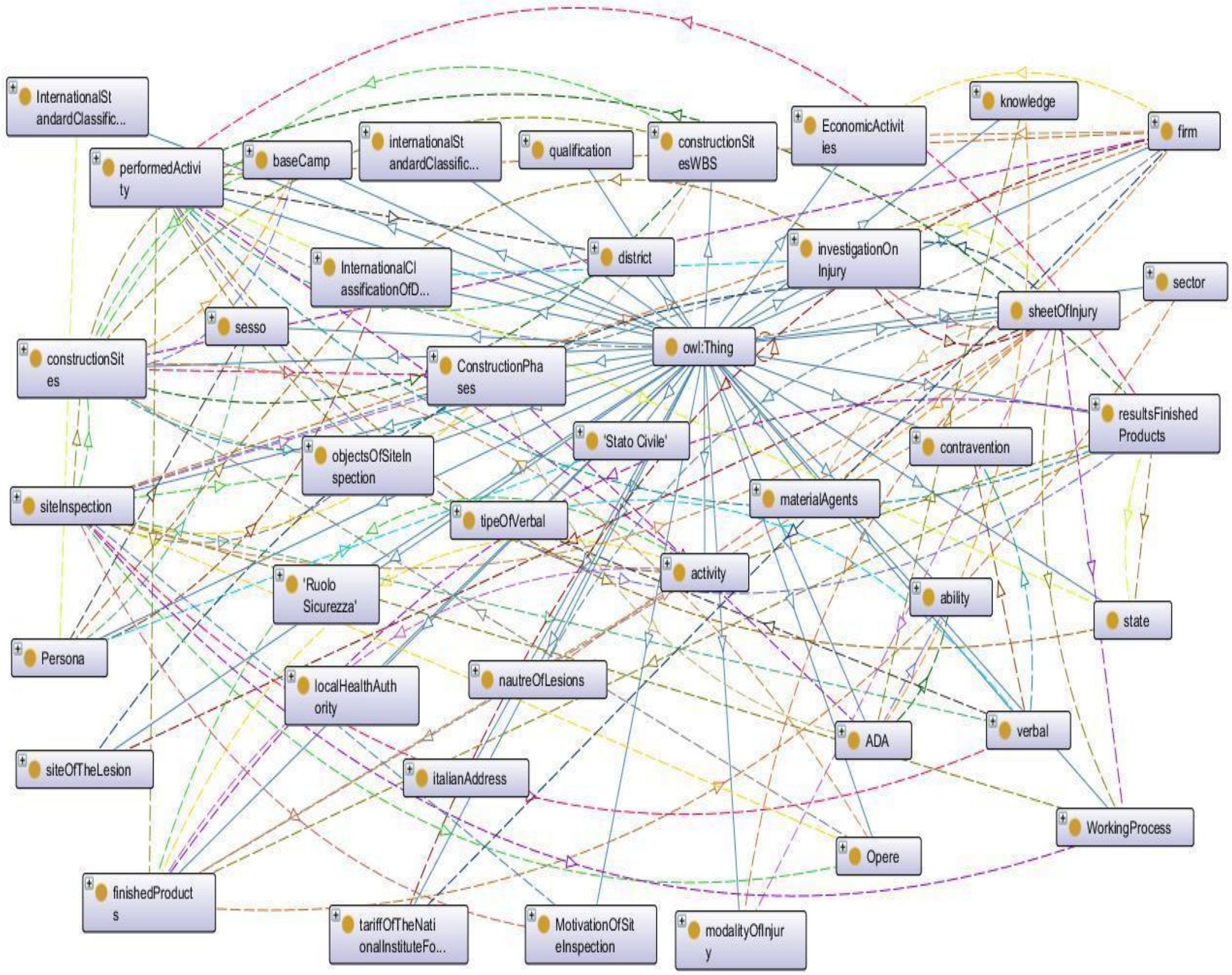
[17] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, X. Wu, Unifying Large Language Models and Knowledge Graphs: A Roadmap, IEEE Trans. Knowl. Data Eng. (2024) 1–20. https://doi.org/10.1109/TKDE.2024.3352100.

[18] A.H. Shariatmadari, S. Guo, S. Srinivasan, A. Zhang, Harnessing the Power of Knowledge Graphs to Enhance LLM Explainability in the BioMedical Domain, (2024).

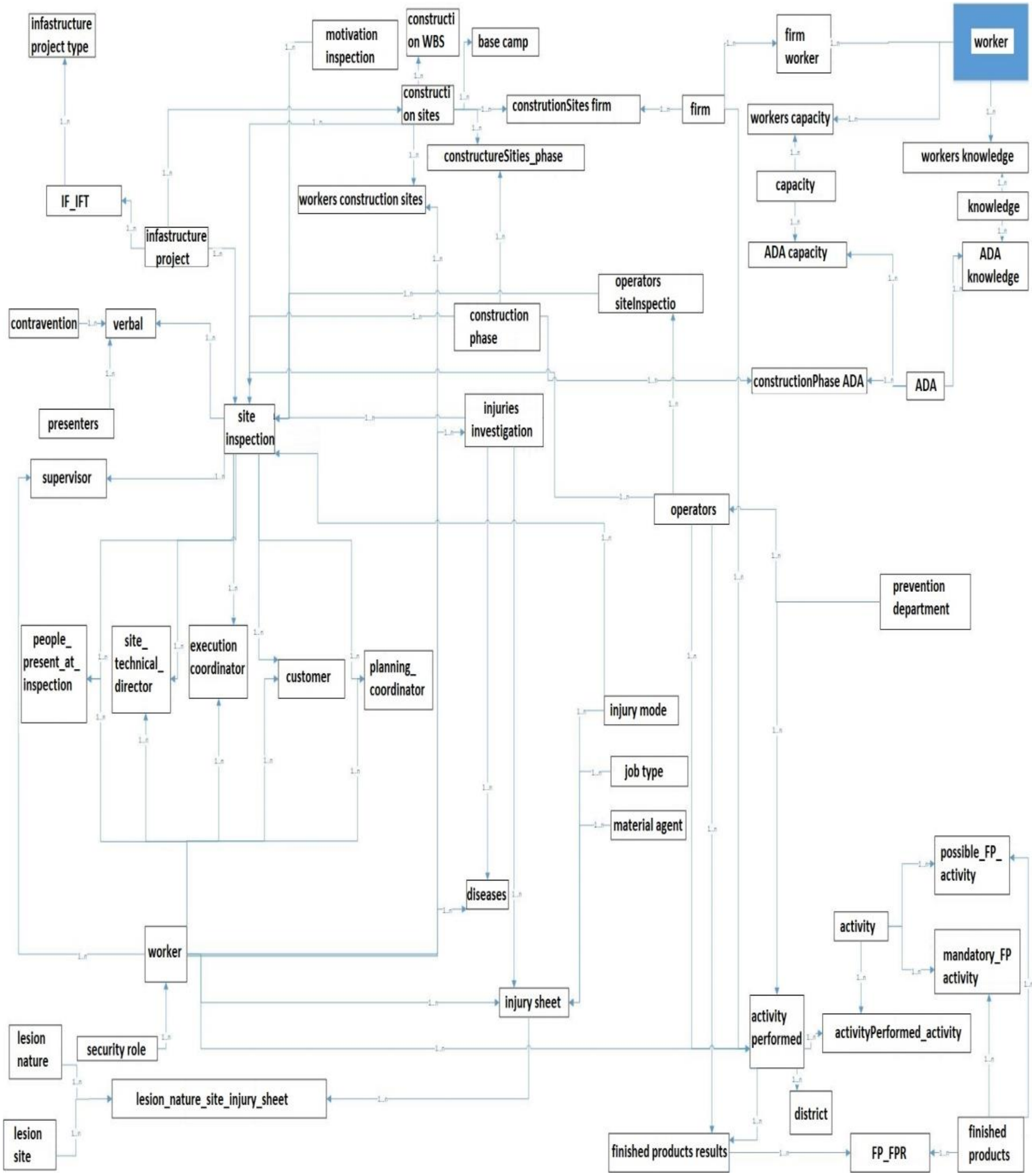
APPENDIX 1: SCHEMA OF THE RELATIONAL DATABASE OF CASE STUDY



APPENDIX 2: ONTOLOGICAL STRUCTURE



APPENDIX 3: NEW DATABASE STRUCTURE


infastructure project type
motivation inspection
constructi on WBS
base camp
firm worker
worker
constructi on sites
construtionSites firm
firm
workers capacity
constructureSities_phase
workers knowledge
IF_IFT
workers construction sites
capacity
knowledge
infastructure project
ADA capacity
ADA knowledge
operators siteInspectio
contravention
verbal
construction phase
constructionPhase ADA
ADA
presenters
site inspection
injuries investigation
supervisor
operators
prevention department
people_ present_at_ inspection
site_ technical_ director
execution coordinator
customer
planning_ coordinator
injury mode
job type
material agent
possible_FP_ activity
diseases
activity
worker
mandatory_FP activity
injury sheet
lesion nature
security role
activity performed
activityPerformed_activity
lesion_nature_site_injury_sheet
district
lesion site
finished products
finished products results
FP_FPR