BUILDING AN ECOSYSTEM FOR UNCOVERING HIDDEN ASSETS: A KNOWLEDGE GRAPH-BASED APPROACH

JOSÉ ALBERTO SOUSA TORRES

DOCTORAL THESIS
IN ELECTRICAL ENGINEERING

DEPARTMENT OF ELECTRICAL ENGINEERING

FACULTY OF TECHNOLOGY

UNIVERSITY OF BRASILIA

University of Brasilia
Faculty of Technology
Department of Electrical Engineering

Building an Ecosystem for Uncovering Hidden Assets: A Knowledge Graph-Based Approach

José Alberto Sousa Torres

DOCTORAL THESIS SUBMITTED TO THE POSTGRADUATE PROGRAM IN ELECTRICAL ENGINEERING AT THE UNIVERSITY OF BRASILIA AS PART OF THE REQUIREMENTS NECESSARY TO OBTAIN THE DEGREE OF DOCTOR.

APPROVED BY:

Georges Daniel Amvame Nze, Ph.D (University of Brasilia)
(Advisor)

Mario Antonio Ribeiro Dantas, Ph.D (University of Southampton)
(External Member)

Geraldo Pereira Rocha Filho, Ph.D (Southwest Bahia State University)
(External Member)

Fábio Lúcio Lopes de Mendonça, Ph.D (University of Brasilia)
(Internal Member)

Brasília/DF, December 2025.

FICHA CATALOGRÁFICA

SOUSA TORRES, JOSÉ ALBERTO

Building an Ecosystem for Uncovering Hidden Assets: A Knowledge Graph-Based Approach. [Brasília/DF] 2025.

209, 210 x 297 mm (ENE/FT/UnB, Doctor, Doctoral Thesis, 2025).

Universidade de Brasília, Faculdade de Tecnologia, Department of Electrical Engineering.

Department of Electrical Engineering

1. Asset Concealment
2. Ontology Engineering
3. Knowledge Graph
4. Fraud
5. Ecosystem
6. Data Integration
7. Entity Matching
8. Record Linkage

I. ENE/FT/UnB
II. Título (série)

REFERÊNCIA BIBLIOGRÁFICA

SOUSA TORRES, JOSÉ ALBERTO (2025). Building an Ecosystem for Uncovering Hidden Assets: A Knowledge Graph-Based Approach. Doctoral Thesis, Publicação PPGEE.217/2025, Department of Electrical Engineering, Universidade de Brasília, Brasília, DF, 192p.

CESSÃO DE DIREITOS

AUTOR: JOSÉ ALBERTO SOUSA TORRES

TÍTULO: Building an Ecosystem for Uncovering Hidden Assets: A Knowledge Graph-Based Approach.

GRAU: Doutor˜ ANO: 2025

JOSÉ ALBERTO SOUSA TORRES

Universidade de Brasília (UnB)

Campus Darcy Ribeiro

Faculdade de Tecnologia - FT

Departamento de Engenharia Elétrica (ENE)

Brasília - DF CEP 70919-970

"Ao meu querido pai, Alberto, in memoriam."

This doctoral thesis is the culmination of a long and challenging journey that would have been impossible to complete alone. I am filled with immense gratitude as I reflect on the many people who have supported, guided, and inspired me along this path.

First and foremost, my most profound appreciation goes to Professor Rafael Timoteo. He was my original guide in this endeavor, and his profound knowledge and mentorship were instrumental in shaping the research's foundational direction. His wisdom and encouragement set me on the right path, and I was deeply saddened when health reasons necessitated his departure from the project. I wish him strength and health. I was incredibly fortunate that Professor Georges stepped in to complete this work. I am profoundly grateful for his willingness to take on this advisory role, for his insightful guidance, and for his steadfast support in maintaining the project's momentum during a critical transition. My sincere thanks also go to the Advocacia-Geral da União. This research would not have been possible without its institutional support and the opportunity to develop it within its walls.

This journey was not mine alone; it was a shared sacrifice, and my most profound gratitude belongs to my family. To my beloved wife, Ane, I have no words to express how much your support has meant to me. You were my anchor. Thank you for your profound patience during my long absences, both physical and mental, as I was lost in this work. Your unwavering belief in me was the only reason I did not give up. To my wonderful son, Pedro, thank you for the joy you brought into my life, even on the most stressful days. I am sorry for the time I missed, and I promise to make up for it. This achievement is as much yours as it is mine. I must also extend my heartfelt thanks to my mother, Lourdes, for her unconditional love and constant prayers, and to my sisters, Tatiana and Daniela, for their unwavering encouragement. Your collective support and belief in me, even from afar, were a constant source of strength.

Finally, I dedicate this work to the memory of the man who inspired it all: my dear father, Alberto. He always dreamed of having a son who would achieve this and one day call him "Doctor."His belief in me was a quiet, constant force long before I began this path. It is my life's most profound sorrow that he passed away before he could hold this finished work in his hands.

Dad, this is for you. I hope I made you proud.

ABSTRACT

Asset concealment and money laundering represent critical challenges worldwide, compromising the integrity of financial systems and governments' ability to recover misappropriated resources. In Brazil, this problem manifests as sophisticated fraud structures that exploit data fragmentation to conceal assets. Although the Brazilian scenario presents particularities, the problem is universal. It stems from the difficulty of connecting heterogeneous databases to identify the ultimate beneficiary of asset slippage schemes, which slows traditional investigations and often makes them ineffective, given the speed at which assets dissipate. This thesis proposes an agnostic computational ecosystem, designed to be adaptable to any jurisdiction or institutional actor involved in combating corruption and financial fraud. The work goes beyond technical implementation, establishing methodological processes for the construction of data-driven ontologies and corporate knowledge graphs, involving essential elements in this context, such as governance and clear role definition, and validating them in a real-world environment with a large volume of data. Furthermore, Artificial Intelligence models focused on retrieving relevant and contextual information were developed, including extracting kinship links with 97% accuracy and creating a social proximity indicator between two people from addresses, without using geocoding. The ecosystem was validated through the construction of an Ontology for Asset Concealment, which created a logical framework that enables the ecosystem to understand and connect fraud patterns in an automated way, and a knowledge graph for asset concealment. The validation was carried out in the Brazilian context, in collaboration between the University of Brasília and the Attorney General's Office, processing a graph with more than 2 billion elements, including entities and relationships, and 15 structuring government databases. The results demonstrated that applying graph construction processes and the developed AI models enabled the detection of concealment networks with multiple levels of depth.

Keywords: Asset Concealment, Crime Investigation, Data Integration, Entity Matching, Knowledge Graph, Ontology Engineering, Record Linkage.

RESUMO

Título: Proposta de Ecossistema para Descoberta de Ativos Ocultados: Uma Abordagem Baseada em Grafos de Conhecimento

A ocultação de ativos e a lavagem de dinheiro representam desafios de escala global, comprometendo a integridade dos sistemas financeiros e a capacidade de governos em todo o mundo de recuperar recursos desviados. No Brasil, essa problemática manifesta-se por meio de estruturas de fraude que se aproveitam da fragmentação de dados para ocultar patrimônio. O problema é universal e está relacionado à dificuldade de conectar bases de dados heterogêneas para identificar o beneficiário final de esquemas de interposição patrimonial, o que torna as investigações tradicionais frequentemente ineficazes. Esta tese propõe um ecossistema agnóstico e adaptável a qualquer ator institucional que atue no combate à fraude financeira. O trabalho estabelece processos metodológicos para a construção de ontologias orientadas a dados e de grafos de conhecimento corporativos, envolvendo elementos essenciais como governança e uma definição clara de papéis, e validando-os em um ambiente real. Foram desenvolvidos, ainda, modelos de Inteligência Artificial para recuperação de relacionamentos ocultos, a exemplo da extração de vínculos de parentesco, com acurácia de 97%, e da criação de um indicador de proximidade entre duas pessoas a partir de endereços, sem o uso de geocoding. O ecossistema foi validado por meio da construção de uma Ontologia para Ocultação de ativos, responsável por criar um arcabouço lógico para o ecossistema, e de um grafo de conhecimento para ocultação de ativos. A validação foi realizada em colaboração entre a Universidade de Brasília e a Advocacia-Geral da União, com a criação de um grafo com mais de 2 bilhões de elementos, entre entidades e relacionamentos, a partir de 15 bases de dados governamentais. Os resultados demonstraram que a aplicação dos processos de construção de grafos e dos modelos de IA desenvolvidos otimizou o processo de detecção de redes de ocultação com múltiplos níveis de profundidade.

Palavras-chave: Correspondência de entidades, Engenharia de Ontologias, Grafos de Conhecimento, Integração de dados, Investigação criminal, Ocultação de ativos, Vinculação de registros.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

AGU	Attorney General's Office of Brazil
AI	Artificial Intelligence
AML	Anti-Money Laundering
ANAC	Brazilian Civil Aviation Agency
ASD	Application Field and Scope Definition Step
BERT	Bidirecional Encoder Representations from Transformers
CADSUS	National Health Card Registration - Brazil
CAFIR	Rural Property Registry
CAGED	General Register of Employed and Unemployed Persons in Brazil
CAR	Environmental Rural Register
CBO	Brazilian Classification of Occupations
CCARCS	Canadian Civil Aircraft Register Computer System
CCIR	Certificate of Registration of Rural Property
CEP	Postal code in Brazil
CGU	Comptroller General of the Union of Brazil
CNAE	National Classification of Economic Activities in Brazil
CNEF	National Address Registry for Statistical Purposes in Brazil
CNPJ	National Registry of Legal Entities in Brazil
CPF	Individual Taxpayer Registry - Brazil
CRF	Conditional Random Fields
DETRAN	State Traffic Departments in Brazil
DGCA	Directorate General of Civil Aviation in India
DGFiP	Système Fiscal de la Direction Générale des Finances Publiques
DIC	Data Cluster Identification Step
DKC	Domain Knowledge Capture Step
DKD	Domain Knowledge Documentation Step
DVLA	Driver and Vehicle Licensing Agency - UK
EDM	Enterprise Data Management
ELP	Entity-Link-Property
ELT	Extract, Load, Transform
EPO	European Patent Office Database
ER	Entity Resolution
ETL	Extract, Transform, Load
FAA	Federal Aviation Administration - US
FCCM	Oracle Financial Crime and Compliance Management
FGTS	Unemployment Fund for Severance Indemnity in Brazil
FHKB	Family History Knowledge Base
FIBO	Financial Industry Business Ontology
FOAF	Friend of a Friend Ontology

GAT	Graph Attention Networks
GCN	Graph Convolutional Networks
GDP	Gross Domestic Product
GDPR	General Data Protection Law
GNNs	Graph Neural Networks
GPT	Generative Pre-trained Transformer
GraphSAGE	Graph Sample and aggregate
GRO—UK	General Register Office Database - UK
HGT	Heterogeneous Graph Transformer
HMM	Hidden Markov Models
IBGE	Brazilian Institute of Geography and Statistics
INSS	Institute of Social Security in Brazil
IMF	International Monetary Fund
IRCC	Immigration, Refugees and Citizenship Canada
IRS	Internal Revenue Service Database US
KBA	Federal Motor Transport Authority - Germany
KG	Knowledge Graphs
LLM	Large Language Model
LSTM	Long Short-Term Memory Networks
ML	Machine Learning
NER	Named Entity Recognition
NHS	National Health Service - UK
NLP	Natural Language Processing
NPD	National Pupil Database - UK
OCO	Ontology Construction
OCR	Optical Character Recognition
OEA	Existing Ontology Alignment Step
OMG	Object Management Group
ORR	Existing Ontology Reusability Review Step
OSA	Optimal String Alignment
OSINT	Open-Source Intelligence
OTK	On-To-Knowledge methodology
OWL	Web Ontology Language
PACER	Public Access to Court Electronic Records in US
PF	Brazilian Federal Police
PIERS	Passport Information Electronic Records System - US
Pje	Electronic Judicial Process - Brazil
RAB	Brazilian Aeronautical Registry
RAIS	Annual Social Information Report of Brazil
RDF	Resource Description Framework
REBEL	Relation Extraction By End-to-end Language generation
RENACH	Brazilian National Driver's License Database
RENAVAM	National Registry of Motor Vehicles in Brazil
REU	Répertoire Électoral Unique - France
RG	Person General Register in Brazil
RL	Record Linkage
RNNs	Recurrent Neural Networks
SENATRAN	National Traffic Secretariat of Brazil
SHACL	Shapes Constraint Language

SISE	Système d'Information sur le Suivi de l'Élève - France
SISOB	Death Control System of Brazil
SNDS	Système National des Données de Santé
SUMO	Suggested Upper Merged Ontology
SUMOF	Suggested Upper Merged Ontology Financial
SVM	Support Vector Machines
SWIFT	Society for Worldwide Interbank Financial Telecommunication
T-KINSHIP	Ontology for Formal Models of Kinship
TED	Descentralized Execution Agreement
TRE	Regional Electoral Courts in Brazil
TSE	Superior Electoral Court of Brazil
TSE-DATABASE	National Voter Registry of Brazil
UDISE+	Unified District Information System for Education Plus India
UK	United Kingdom
UML	Unified Modeling Language
UnB	University of Brasilia
UNODOC	United Nations Office on Drugs and Crime
USA	United States of America
USCG	United States Coast Guard
USPTO	United States Patent and Trademark Office Database
VAHAN	National Register of Motor Vehicles - India

CHAPTER 1

INTRODUCTION

Asset concealment fraud represents one of the most persistent threats to global economic systems, with impacts that transcend national borders and financial sectors. Multilateral bodies such as the United Nations Office on Drugs and Crime (UNODC) and the International Monetary Fund (IMF) estimate the annual global value of laundered funds, or hidden assets, to be between 2% and 5% of global Gross Domestic Product (GDP) (INTER-AMERICAN DEVELOPMENT BANK, 2024). This range corresponds to a staggering annual monetary value of approximately US$800 billion to US$2 trillion.

The consequences of this massive concealment are profound and multifaceted. It distorts markets, inflates asset prices, and undermines the integrity of the global financial system. It also erodes trust in public institutions, finances the expansion of other criminal activities, and exacerbates social inequality by allowing criminal and corrupt elites to operate outside the rule of law while draining essential public resources (TORRES *et al.*, 2024).

Simser describes some of the most common methods for concealing assets (SIMSER, 2008), including transferring ownership to trusted individuals, such as family members or close associates. Complex domestic and offshore corporate structures can also obscure the true ownership of assets, while trusts can conceal the true owner and create investigative difficulties due to confidentiality restrictions.

Recovering stolen assets is a complex process that requires cooperation between domestic agencies and ministries in multiple jurisdictions, each with its own legal systems and procedures (BRUN *et al.*, 2021). Governments must work together to close the gaps that facilitate asset concealment, including enforcing strict reporting requirements, improving the transparency of financial transactions, and promoting information sharing across jurisdictions.

Asset recovery investigations require large amounts of data from the public and private sectors, such as tax records and bank transactions, which track monetary flows and can reveal

lifestyle details. Investigators typically use two approaches: follow-the-money, aimed at locating specific funds, and net-worth investigations, which seek to identify all assets owned by an individual (KENNEDY, 2007). When examining such cases, they must combine fragmented data to create a unified asset history. However, they encounter difficulties because essential data is spread across multiple sources and organized into distinct compartments, making it challenging to discover relationships.

This global phenomenon of financial opacity finds a critical expression in the Brazilian public administration. As a major emerging economy with a complex tax and corporate legal framework, Brazil faces persistent challenges in recovering public funds diverted through sophisticated fraud. The bridge between the global challenge and the Brazilian reality lies in the structural asymmetry of information. While fraudsters operate with high fluidity and speed across borders and sectors, the State's response is often slow and compartmentalized, turning the Brazilian case into a representative microcosm of the global fight against financial crime.

In Brazil, fraud involving the concealment of assets is a common obstacle to the effectiveness of tax enforcement and to the reimbursement of the public treasury. Debtors often deliberately use this scheme to deplete their assets and frustrate the fulfillment of financial obligations definitively established in favor of the federal government. The Attorney General's Office (AGU) is the judicial representation of the Brazilian State. It continually faces the challenges posed by such strategies, which utilize complex business and corporate structures to conceal the ownership of assets and rights.

AGU has access to many government databases that cover registration, tax, corporate, and asset information. However, the mere availability of these information repositories does not, per se, translate into optimized investigative capacity to detect complex asset concealment schemes. There is a lack of a unified technological and methodological strategy for the large-scale processing, cross-referencing, and integrated analysis of this data. This operational gap limits the agency's proactivity and gives fraudsters an advantage due to the fragmentation of government information.

To address these challenges, it is necessary to distinguish between two distinct orders of problems that this thesis aims to bridge. Current literature in Knowledge Engineering and Computer Science lacks formal semantic models (ontologies) specifically designed for the do-

main of asset concealment. Existing financial ontologies are often too generic or focused on banking transactions, failing to capture the nuances of fraudulent intent, kinship-based "straw man"relationships, or the temporal evolution of asset transfers. Furthermore, there is a lack of data-driven methodologies for constructing Knowledge Graphs (KGs) in highly adversarial and heterogeneous environments where unique identifiers are frequently missing.

Additionally, in the Brazilian context, the Attorney General's Office (AGU) possesses access to vast governmental databases. However, there is no unified technological strategy to cross-reference this data. The operational gap is not a lack of data, but the absence of a tool capable of performing integrated large-scale analysis. Currently, investigations are often manual or localized, giving a strategic advantage to offenders who exploit the lack of interoperability between registration, tax, and corporate repositories.

In light of the above, this thesis proposes the design and development of a computational ecosystem for integrating government databases, specifically to facilitate the investigation of asset concealment. Unlike a platform, which provides a static environment for execution, or an architecture, which defines the structural components, the ecosystem proposed here is a dynamic, synergistic framework. It encompasses a multi-layered technological architecture based on Knowledge Graphs; a set of evolving methodological processes for knowledge modeling; and adaptive AI models for relationship discovery. It is an "ecosystem"because its components are interdependent: the ontology informs the graph, the graph feeds the AI, and the AI's results refine the ontology, allowing the system to adapt to new fraud patterns.

The proposed ecosystem was designed with a level of abstraction that allows for its potential use by other Brazilian public administration bodies, such as the Public Prosecutor's Office, the Audit Courts, and the judicial police, which face similar challenges. The conceptual and technological structure is sufficiently agnostic to be adapted to international contexts. Governments in other countries can implement it with an equivalent public data framework, representing a broader contribution to the fight against fraud and corruption.

The originality of this work lies in the semantic-proactive approach. While existing tools focus on reactive search, this proposal introduces a proactive detection model through: the first formal ontology explicitly dedicated to the patterns of asset concealment; original AI models for inferring kinship and social proximity in the absence of unique keys; a replicable methodology

for KG construction that integrates business experts into the data engineering loop.

1.1 DELIMITATION AND SCOPE

To ensure the feasibility and focus of this research, the following delimitations are established:

- Non-Scope (Out of Scope): This thesis does not cover the automatic execution of asset seizures, legal adjudication of guilt, or the creation of new primary data sources. It is focused on the investigative intelligence phase.
- Legal and Ethical Limits: The research is strictly compliant with the Brazilian General Data Protection Law (LGPD). The proposed models do not perform automated profiling with legal consequences without human intervention; rather, they provide ranked leads for human investigators.
- Operational Limits: The prototype validation is limited to the databases available via the technical-scientific cooperation between UnB and AGU (TED).
- International Applicability: While the primary use case is Brazilian, the methodologies and the ontology are designed to be agnostic, allowing for implementation in any jurisdiction that follows standard multilateral financial transparency guidelines.

Given the above, this thesis seeks to answer the following research question: How can the systemic and semantic integration of fragmented government databases enhance the Attorney General's Office's analytical capacity to proactively detect asset concealment fraud? The hypothesis is that creating a computational ecosystem to integrate heterogeneous data sources will significantly increase state agencies' analytical capacity by transforming isolated data points into actionable intelligence through semantic correlation.

In this way, overcoming information fragmentation through a modular, interoperable architecture will enable the generation of actionable intelligence by correlating previously isolated information, making the investigative process more efficient, proactive, and assertive. As detailed in the subsequent chapters, this hypothesis will be validated through the proposed ecosystem's design, development, and evaluation.

1.2 OBJECTIVES

The main objective of this work is to propose an agnostic ecosystem for integrating and analyzing data on asset concealment and recovery, thereby improving the accuracy and efficiency of investigations.

For this, the following specific objectives can be listed:

- Propose a process for constructing domain ontologies, prioritizing data analysis and validation in real-world applications.
- Develop a domain ontology for data mapping and integration in asset recovery.
- Systematize and propose a methodology for constructing domain Knowledge Graphs in a corporate context.
- Develop and evaluate Artificial Intelligence models to discover implicit relationships in government data, specifically focusing on inferring kinship ties and measuring social proximity.

1.3 CONTRIBUTIONS

This thesis presents contributions beyond implementing a technological solution by proposing reusable methodologies and creating artifacts directly impacting computer science, knowledge engineering, and the specific domain of financial investigation. These contributions have been categorized into methodological, artifact, and applied research contributions.

In the methodological scope, the thesis proposes a process for constructing domain ontologies. This process differs from the conceptual approaches, characterized by a data-driven orientation and a focus on integration from the initial phases. The emphasis on identifying data clusters and validating them in real-world applications gives the methodology a pragmatic, scalable character, especially suitable for highly complex domains, such as asset investigation. One of the artifact contributions is the development of the first formal ontology for asset concealment. It fills a significant gap in the literature by providing a standardized, reusable semantic schema that can underpin future research and systems to combat financial crimes.

This thesis advances beyond existing methodologies by detailing an improved process for constructing Knowledge Graphs (KGs). This approach innovates by integrating the active participation of business experts, data governance, and the continuous refinement of the ontology as supporting activities. The proposed process addresses practical gaps in corporate environments, including managing business requirements, ensuring data quality, and providing compliance assurance.

The applied research contribution is consolidated through the design, implementation, and validation of a computational ecosystem in a real-world scenario, in collaboration with the Brazilian Attorney General's Office, which demonstrated the solution's effectiveness in a complex context of significant importance to the Brazilian government.

This work also contributes to the technical and applied levels when proposing solutions to specific data integration challenges. In this regard, artificial intelligence models were developed to discover hidden relationships, such as kinship ties, from data without unique identifiers, a heuristic for detecting non-trivial links between environmental offenders, and social proximity based on address information. These are technical solutions to recurring, complex problems in integrating government data that can be adapted to other contexts.

In summary, this thesis's central contribution lies in conceiving, developing, and validating a complete socio-technical framework. This framework encompasses methodological processes, technological artifacts (notably the domain ontology and the computational ecosystem), and applied AI models to solve data fragmentation in the fight against asset concealment. Validation in a highly relevant practical context attests to the value and originality of the research.

1.4 PUBLISHED PAPERS

- TORRES, José Alberto Sousa et al. Using spatial data and cluster analysis to automatically detect non-trivial relationships between environmental transgressors. In: 2022 IEEE International Conference on Data Mining Workshops (ICDMW). IEEE, 2022. p. 98-104.

- TORRES, José Alberto Sousa et al. Ontology Development for Asset Concealment Investigation: A Methodological Approach and Case Study in Asset Recovery. Applied Sciences, v. 14, n. 21, p. 9654, 2024.

1.5 THESIS ORGANIZATION

This thesis is organized into nine chapters, each covering an essential aspect of the research, from the conceptual model to the implementation and evaluation of the proposed ecosystem.

- Chapter 1 — Introduction: This chapter provides an overview of the study's research problem, objectives, and significance. It introduces key concepts and the motivation behind using ontologies and knowledge graphs in asset concealment investigations.
- Chapter 2 – Literature Review: Presents existing literature on asset concealment, investigative techniques, and tools, including ontologies, knowledge graphs, entity matching, record linkage methods, and asset concealment investigation tools like OSINT, semantic analysis, and financial transaction analysis.
- Chapter 3 — Conceptual Framework: This chapter defines the conceptual model used in the study, covering elements like data, processes, AI models, and outcomes that structure the proposed ecosystem.
- Chapter 4 – Proposal for an Ontology Building Process: Outlines a data-driven process for ontology development, including specification, implementation, and evaluation stages to ensure efficient representation of asset concealment-related knowledge.
- Chapter 5 – Proposal for a Knowledge Graph Building Process: Proposes a methodology for building domain knowledge graphs, including business understanding, data acquisition, knowledge improvement, visualization, and evaluation to enhance investigative capabilities.
- Chapter 6 – Proposal for an Asset Concealment Ontology: Describes the development of the specific ontology for asset concealment investigation, applying the process detailed in Chapter 4.
- Chapter 7 — Ecosystem Implementation: This chapter details the practical implementation and validation of the proposed ecosystem. It follows the knowledge graph building process, covering business understanding, environment planning, knowledge acquisition, and knowledge improvement with real-world data from a Brazilian public agency.

- Chapter 8 – Results and Discussion: Presents and discusses the results obtained from the proposals detailed in Chapters 4 (Ontology Building Process), 5 (Knowledge Graph Building Process), and 6 (Asset Concealment Ontology), evaluating their effectiveness.

- Chapter 9 – Conclusion: Summarizes the key findings of the thesis, highlights its main contributions to the field, and suggests future research directions, emphasizing the impact of the proposed ecosystem in financial crime investigations

CHAPTER 2

LITERATURE REVIEW

This chapter presents a literature review and establishes the foundations for the proposed ecosystem. It begins by contextualizing the complexities of the asset concealment process and analyzing asset recovery frameworks to highlight investigators' challenges. Next, it addresses the inadequacy of relational databases for modeling the complex networks of entities and relationships involved in these illicit schemes, paving the way for discussing more flexible, semantic architectures, such as ontologies and KGs.

The chapter delves into the methods and techniques for building asset concealment frameworks. It examines methodologies for extracting entities from unstructured sources, such as Named Entity Recognition (NER), and unifying identities across datasets through Record Linkage. We also analyzed the problem using graph neural networks (GNNs) and large language models (LLMs). Based on this, we review the tools used in the field, categorized by function, from data integration and Open Source Intelligence (OSINT) to link analysis and Anti-Money Laundering (AML) platforms. This technology landscape analysis demonstrates that, while powerful components exist, they often operate in isolation.

We also examined existing processes for building ontologies and Knowledge Graphs to identify methodological gaps. This review culminates in articulating the shortcomings of the literature, thereby justifying the need for this thesis's integrated, multidisciplinary approach to overcome the persistent challenge of data fragmentation in asset concealment investigations.

The field of data science has produced many essential components that can be used to address part of the problem of asset concealment investigations (methodologies, techniques, tools). However, it has failed to assemble them into a coherent, governable, and legally defensible ecosystem, specifically tailored to the unique constraints of the public sector. The literature presents solutions for parts of the problem—entity extraction, graph storage, network visualization—but lacks an end-to-end architectural and methodological design. It is precisely this

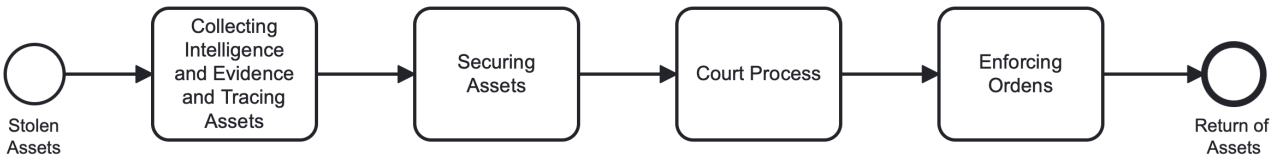

Figure 2.1: Asset Recovery Process (BRUN *et al.*, 2021)

gap that the proposed ecosystem aims to fill. This line of reasoning should become the guiding thread of the entire chapter.

2.1 ASSET CONCEALMENT

The progress made in the last few decades in asset concealment and recovery, particularly in the fight against corruption, is widely recognized as necessary. Many steps in the asset recovery process have been intensely studied and documented. However, different asset detection and recovery phases have challenges that must be overcome (UNODC, 2023).

(BRUN *et al.*, 2021) proposes one of the best-known processes for Recovering Stolen Assets (Figure 2.1). It consists of five main steps. Collecting Intelligence and Evidence and Tracing Assets is the step in which evidence is gathered from open-source intelligence and governmental and private databases to identify and trace assets. The next step outlines the actions to prevent the dissipation, movement, or destruction of the assets identified during the investigation that are subject to confiscation. Criminal conviction, asset confiscation, or private civil actions may be executed during the court process. After that, measures must be taken to ensure that the restraint, seizure, or confiscation of assets orders are carried out. At the end of the process, the result is the return of the assets to the requesting jurisdiction by application.

As discussed, many of the processes' challenges are technical, legal, or even international cooperation-related. However, this work will focus solely on analyzing and proposing solutions to technical issues in the Collecting Intelligence and Evidence and Tracing Assets stage. One of the significant difficulties in asset recovery cases is providing evidence that connects the assets to criminal activity or proving the value of the benefits gained from the crime. To establish this connection, practitioners need to locate and trace assets until they can link them to the crime or find them. However, it is common for these assets to have been moved globally through complex schemes involving offshore accounts, corporate entities, nominees, intermediaries, "straw men,"and various financial transactions to launder the money and obscure the connection to

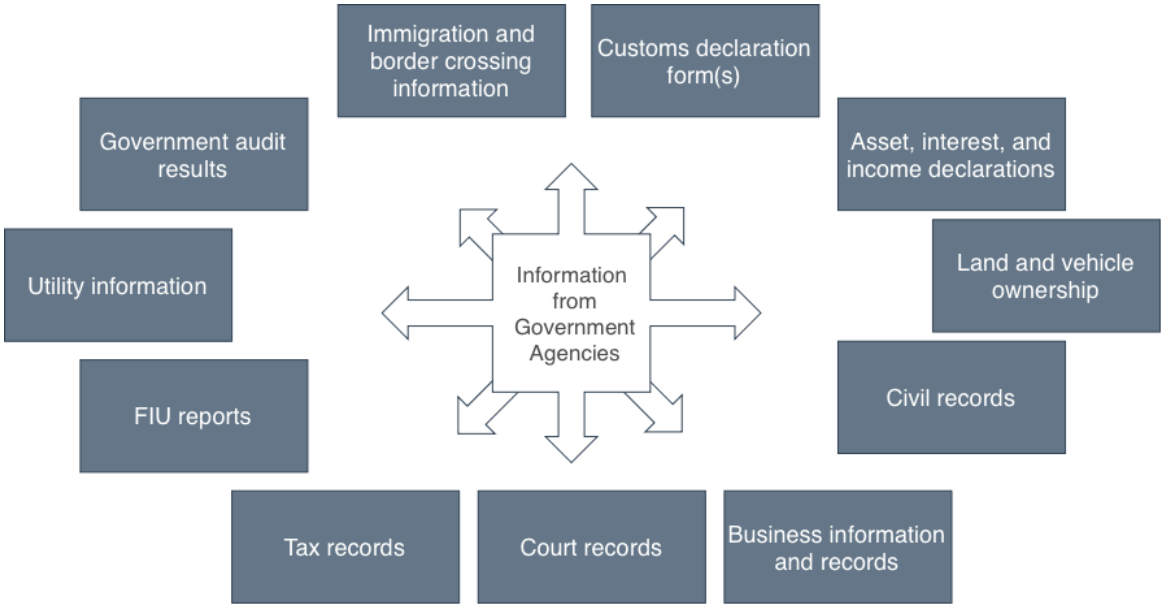

Figure 2.2: Data Sources Often Used in Asset Concealment Investigations (BRUN *et al.*, 2021)

the crime.

According to (BRUN *et al.*, 2021), asset recovery cases are often document-intensive, time-consuming, complicated, and demanding of practitioners need to be able to obtain relevant information through traditional investigative techniques, comprehensively use the information available through domestic and foreign open sources, collect contextual or circumstantial evidence; understand what information can be obtained from financial institutions; analyze bank statements, business records, financial documents, and contracts; determine the ultimate beneficial owners; and organize the information in a timely, comprehensive, and coherent manner. In this way, some questions need to be answered, such as where the assets are now and how they were moved. Aiming to help answer these questions and solve these challenges, the World Bank listed some sources of preliminary Information from Government agencies, as presented in Figure 2.2 (BRUN *et al.*, 2021).

However, even though the origin of the process, i.e., examples of data sources, is partially mapped, there is a gap in the entire remaining process until the final location of the asset is obtained and its current links and possible movement history are discovered. Although several tools help with data processing, visualization, and analysis, as presented in the following sections, the main problem that persists is identifying the physical data sources, not the categories, and going all the way until the available tools can use them to provide value.

2.2 METHODS, TECHNIQUES, AND PROCESSES USED IN ASSET CONCEALMENT INVESTIGATION

Asset concealment investigations often use several methods, techniques, and processes. Because traditional databases with strictly relational schemas offer limited support for representing the intricate and evolving interconnections among entities, modern investigative strategies have shifted toward more flexible, semantically rich architectures, including ontologies and Knowledge Graphs (KGs). These enable dynamic, context-aware analyses of suspect entities and their relationships.

Defining the domain through entities and relationships solves only part of the problem. Identifying these elements in the data to instantiate the ontology and build the KG is just as complex. It is common for entities to be hidden in unstructured data (e.g., names, addresses, personal identifiers), or for the same entity to be replicated across multiple databases. A standard solution to these two problems is to create specific models using techniques such as Named Entity Recognition and Record Linkage. We also evaluated the use of graph neural networks (GNNs) and large language models (LLMs) for problem analysis.

2.2.1 Knowledge Engineering

Knowledge Engineering is formally defined as the engineering discipline that develops Knowledge-Based Systems. Initially, Knowledge Engineering was seen as equivalent to the transfer of knowledge from an expert to a knowledge base; however, approaches based on this perspective often failed because experts were usually unaware of the experiences they used to solve their problems (STUDER *et al.*, 1999). This observation led to the development of various modeling frameworks, in which building a knowledge-based system means creating a computational model that achieves problem-solving capabilities comparable to those of a domain expert (SCHREIBER, 2008).

In the 1990s, the attention of knowledge engineering gradually shifted to domain knowledge, particularly to reusable representations in the form of ontologies (SCHREIBER, 2008). During this decade, ontologies gained widespread attention as vehicles for sharing concepts across

distributed communities, such as the web. Ontology engineering, in turn, is the discipline that deals with the construction and maintenance of ontologies, providing guidelines for their construction (SCHREIBER, 2008).

More recently, a new element has emerged in knowledge engineering: the Knowledge Graph. Popularized by large technology companies like Google, knowledge graphs represent an evolution of traditional knowledge bases, focusing on data interconnection to create a network of contextually rich information (SINGHAL, 2012).

A knowledge graph models a network of real-world entities (such as people, organizations, products, or concepts) and the relationships between them. Unlike a relational database, which stores data in rigid tables, a knowledge graph uses a graph structure composed of nodes (entities) and edges (relationships). This flexible structure is ideal for representing complex, heterogeneous data, enabling the discovery of connections and patterns that would otherwise be hard to see. Ontology, developed in the earlier phase of knowledge engineering, often serves as the blueprint for the knowledge graph, defining the types of entities and relationships that can exist. As with ontologies, specific studies are developing processes for constructing knowledge graphs.

2.2.1.1 Ontologies

In the context of information science and artificial intelligence, the term ontology was first defined as an "explicit specification of a conceptualization" (GRUBER, 1993). This definition was later changed by (BORST, 1997) as a "formal specification of a shared conceptualization." It added to the initial definition of the concept that the conceptualization should express a shared view between parties and can be expressed in a machine-readable format (GUARINO *et al.*, 2009).

In this context, "the formalization of knowledge in declarative form begins with a conceptualization and includes objects presumed or hypothesized to exist in the world and their relationship" (GENESERETH; NILSSON, 1987). Objects can be physical, such as Vehicles or People, or abstract, such as Family. In turn, the formal specification concept is related to the need to use a language to represent objects and their relationships. The central part of

an ontology consists of a hierarchy of these objects and the relations among them. Unary and binary predicates, respectively, represent objects and relations.

When mapping collective knowledge of terms used within a specific community, the terminology is generally predetermined by all members, leading to a shared understanding of those terms. It can often lead to terms being defined imprecisely or in a way that allows multiple interpretations. This situation can be challenging because domain ontologies from different communities must be integrated. It is a widely adopted practice to employ a foundational ontology to mitigate terminological and conceptual uncertainties that may arise during integration and to address this challenge (OBERLE *et al.*, 2007) efficiently.

To the best of our knowledge, no ontology or knowledge graph has yet been designed to map knowledge related to the Asset Concealment domain. However, in top-level ontologies, some concepts and relationships associated with this specific domain were shared with other domains, such as finance and family relationships (TORRES *et al.*, 2024). A fundamental ontology, or upper-level or top-level ontology, constitutes an axiomatic framework concerning overarching domain-independent classifications within reality, encompassing entities, properties, relationships, etc. It is often used as an initial stage for developing novel domain-specific ontologies. Rather than starting the modeling process from the ground up, a foundational ontology provides a predetermined set of ontological elements that can be used to construct these ontologies.

In this regard, a survey of potential top-level ontologies relevant to these domains was conducted to enable the reuse of concepts in the proposed asset concealment ontology.

A) Financial Domain

We selected three ontologies of the financial domain to analyze: the Financial Industry Business Ontology (FIBO) (BENNETT, 2013), the Suggested Upper Merged Ontology (SUMO) (NILES; PEASE, 2001) with its specific domain ontology SUMO Financial (SUMOF), and the ONTO-FIC ontology (ABROUK *et al.*, 2023).

The Financial Industry Business Ontology is a standardized model created by the Enterprise Data Management (EDM) Council in collaboration with the Object Management Group (OMG) (BENNETT, 2013). It formalizes the concepts and relationships that define the structure

and meaning of terms widely used across the financial industry (TORRES *et al.*, 2024). The core purpose of FIBO is to provide a unified language and structure for financial institutions to represent their data, which is essential for regulatory compliance, risk management, and operational efficiency (TORRES *et al.*, 2024). FIBO's interoperability enables systems within and between organizations to communicate more effectively, thanks to a common framework that facilitates data exchange and understanding. This feature is particularly beneficial for financial institutions that must accurately and promptly aggregate and analyze data for risk management purposes (TORRES *et al.*, 2024).

FIBO is structured into ten ontologies that encompass various aspects of the financial industry. These ontologies belong to the following domains: FIBO Business Entities, Business Processes, Corporate Actions and Events, Financial Business and Commerce, Foundations, FIBO Indices and Indicators, FIBO Loan, Market Data, and FIBO Securities. In addition to the base ontology (Foundations), many others have concepts related to the domain under analysis, such as the Financial Business and Commerce and FIBO Business Entities ontologies (TORRES *et al.*, 2024).

The Suggested Upper Merged Ontology is a comprehensive and formal ontology that is a foundation for building domain-specific ontologies (NILES; PEASE, 2001). Within SUMO, various domains, including finance, can be represented and structured semantically richly. SUMO Financial explicitly addresses the concepts, relationships, and operations that define the financial domain, allowing for a more precise and standardized representation of financial data, processes, and entities (TORRES *et al.*, 2024). By incorporating SUMO Financial, it is possible to create a unified understanding of financial concepts such as assets, liabilities, transactions, and economic events. SumoF can represent the payment for a financial transaction or the transfer of funds, assets, or services in the financial domain (TORRES *et al.*, 2024). It conceptualizes several financial domains, including deposits, payments, withdrawals, investments, real estate, and pension plans.

Onto-FIC is a domain ontology for interbank transactions in the Society for Worldwide Interbank Financial Telecommunication (SWIFT) network. It was built on extracting knowledge from financial articles to populate a "Know Your Customer"ontology for banks (ABROUK *et al.*, 2023). This ontology enables the modeling of SWIFT transactions and clients. It also has

a fraud-detection approach that complements the tools organizations use to prevent and detect financial fraud. The concepts defined in this ontology are organized into three categories: the customer, the financial institution (which is an agent), and the financial transactions (TORRES *et al.*, 2024). The complete set of Onto-FIC concepts is presented in the table 2.1

Table 2.1: Onto-FIC Concepts (ABROUK *et al.*, 2023).

Category	**Description**
Activity	The domain of activity
Agent	Financial institution (customer's bank, intermediary bank)
Agent:CreditorAgent	Financial institution associated with the creditor customer
Agent:DebtorAgent	Financial institution associated with the debtor customer
Agent:IntermediaryAgent	Financial institution acting as an intermediary between two financial institutions
Customer	Creditor and debtor customer (not a financial institution
Customer:Organization	Entity engaged in selling goods and services
Customer:Organization:Association	Organizations, associations, and NGOs. Not used due to legal ambiguity.
Customer:Organization:Company	Public or private company
Customer:Person	Individual
Transaction	Money transfer between two customers

B) Kinship and Friendship Domains

When discussing asset concealment, relationships between people, especially kinship and friendship relationships, are essential for analyzing the problem (TORRES *et al.*, 2024). An extensive search for related work was conducted to identify the main ontologies in these domains. In this context, we will analyze a domain ontology about genealogy, including OntoUML (CARVALHO *et al.*, 2014), the Family History Knowledge Base (FHKB) (STEVENS *et al.*, 2014), the Kinship Ontology (CHUI *et al.*, 2020), and the Friend of a Friend Ontology (FOAF) (BRICKLEY; MILLER, 2014).

OntoUML is based on the Unified Modeling Language (UML) class diagram and distinguishes class categories according to UML's type taxonomy (CARVALHO *et al.*, 2014). The

domain ontology about genealogy in OntoUML tries to define relationships between biological ancestors (TORRES *et al.*, 2024). "It is based on the stance that human beings (Person) are products of instantaneous Conception events, which occur when a human male (MaleProcreator) sperm unites with a human female (FemaleProcreator) oocyte egg"(CARVALHO *et al.*, 2014). The concepts found in the Ontology are presented in Figure 2.3

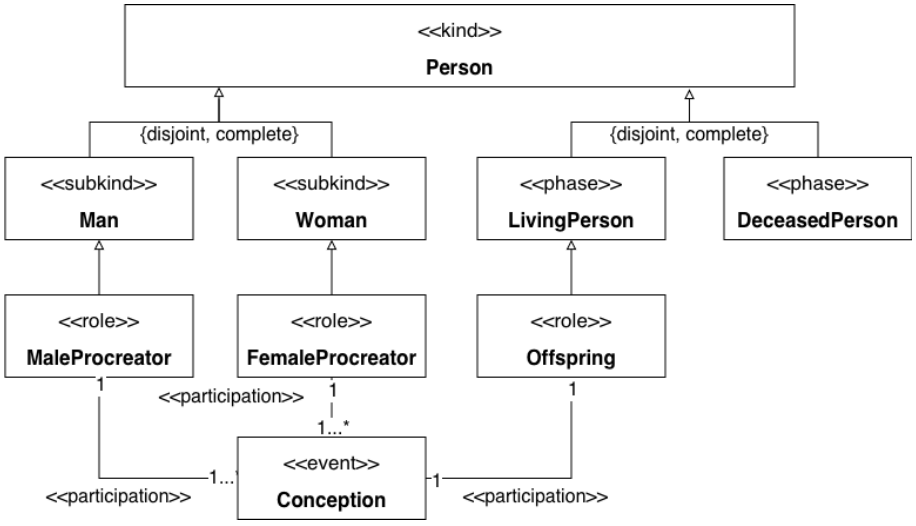

Figure 2.3: Genealogy ontology based on OntoUML. (CARVALHO *et al.*, 2014)

Family History Knowledge Base is based on ten core classes: DomainEntity, with three subclasses Person, Sex, and Partnership; Sex, with Male and Female as subclasses; Man and Woman, which are defined as a Person that has a hasSex property to relate this with some Male or Female classes; and Partnership with its subclass Marriage (STEVENS *et al.*, 2014). FHKB also uses properties to describe required information about an individual, including their parentage and siblings (TORRES *et al.*, 2024).

Chui proposes an Ontology for Formal Models of Kinship (T-KINSHIP) using a first-order ontology that captures anthropological concepts of kinship (TORRES *et al.*, 2024). Her approach differs from existing work because "instead of focusing on which kinship relationship types should be considered as roles, she has taken the existing anthropological kinship terms and has axiomatized them as binary relations within the ontology"(CHUI *et al.*, 2020). According to this work, finding a path between two entities is necessary to determine a relationship between two people in the graph. In this way, all kinship relations begin with the parent/child relation, and all other relations are defined through composition.

Friend of a Friend ontology is a project that aims to link people and information using

the Web. It tries to integrate social networks of human collaboration, such as friendship and association (BRICKLEY; MILLER, 2014). FOAF enables the creation of social networks that can be used to define relationships among customers, bank personnel, and roles and hierarchies within organizations (KUL; UPADHYAYA, 2015). Its vocabulary Specification defines essential classes, such as Group, which represents a collection of individual agents; Person, which means people and is a subclass of Agent; and Organization, which corresponds to social institutions such as companies and societies and is also a subclass of Agent(BRICKLEY; MILLER, 2014). This ontology's two most important relationships are "knows"and "member"(TORRES *et al.*, 2024). The first defines a person known by this person and indicates some reciprocated interaction. This relationship does not imply friendship. In addition, a "member"relationship indicates that a person is a member of a Group.

2.2.1.2 Knowledge Graphs and Context-Aware Computing

Knowledge Graphs represent knowledge and serve as the foundation for various industrial applications. The interest in this technology stems from its structure, which supports domain conceptualization and data management, and from its role as a key driver of various Artificial Intelligence applications. The KG presents a set of real-world entities linked by semantically related relationships. Information is assigned formal semantics through data annotation and manipulation in a machine-readable format, thereby reducing ambiguity (DEPREZ *et al.*, 2025).

The term Knowledge Graph emerged in the second decade of the 21st century, following Google's launch of the concept (SINGHAL, 2012). According to Google, this structure helps efficiently locate accurate information by understanding the connections between things and using contextual details to prevent word ambiguity. Furthermore, KGs would also offer a brief overview of pertinent information related to a particular query, including essential facts about a specific entity, and assist in uncovering unexpected findings by proposing related searches.

Despite Google presenting it as a new development, it is essential to recognize that the concepts behind KGs emerged from graph theory and semantic networks. From a technical perspective, a Knowledge Graph is a labeled multi-graph comprising entities (nodes or vertices) and relationships (edges, facts, or links) that connect these entities. It concerns the accumula-

tion of factual knowledge stored as triples and provides a comprehensive representation through typed relationships. Because of that, it is often used to depict intricate information.

The semantics of the KG are provided by an ontology, along with a reasoning engine to detect violations of the ontology, making it a mature approach for expressing and manipulating domain knowledge (KEJRIWAL, 2022). The direction of the link between nodes in a KG is significant as relations are not necessarily symmetric, making it a directed graph where the head entities point to the tail entities via the relation's edge (PENG *et al.*, 2023).

A knowledge graph can be populated through human-driven, semi-automated, or fully automated approaches, guaranteeing that the information stored is easily comprehensible and confirmable by individuals. Querying a knowledge graph entails navigating it, with tasks reduced to traversing the graph to fetch specific information, which is beneficial across different problem categories (CHAUDHRI *et al.*, 2022).

KGs are also used as enablers for knowledge management in context-aware systems. They act as the fusion point for heterogeneous data and serve as the system's memory, allowing for the retrieval of contextualized content (OLARU *et al.*, 2011). Ontology-based context models - similar to the architecture proposed in this thesis - are widely used to enable semantic reasoning and to represent complex context relationships through knowledge graphs. The graph structure is ideal for capturing the "multirelational interactions between users and services in different contexts"(e.g., time, location, social profiles) (OLARU *et al.*, 2011).

Context-Aware Computing is an established computational paradigm in which systems can discover and take advantage of contextual information (ABOWD *et al.*, 1999). Instead of operating based on explicit user input, context-aware systems dynamically adapt their behavior based on the environment, with context being any information that can be used to characterize the situation of an entity (ABOWD *et al.*, 1999). Asset concealment can be understood as a practice of context obscuring. Its operational logic consists of breaking or distorting the links between individuals and assets to make the ownership relationship invisible or ambiguous. This process exploits the inability of traditional audit systems, often based on isolated data sources, to reconstruct the dispersed informational nexus.

No public knowledge graph related to the area under study has been identified. We can also not infer whether national governments use a similar structure in their private research

activities.

2.2.1.3 Processes to Build Ontologies

Building ontologies is a critical process in knowledge representation and semantic technologies. Its construction typically involves several phases, from identifying relevant concepts and relationships to their formal representation using standardized languages such as the Resource Description Framework (RDF) or the Web Ontology Language (OWL) (TORRES *et al.*, 2024). This process ensures semantic consistency, interoperability, and reusability across systems. This section will explore some methodological processes for developing ontologies, as described in (IQBAL *et al.*, 2013), (FERNÁNDEZ-LÓPEZ; GÓMEZ-PÉREZ, 2002), (ELHASSOUNI; QADI, 2022), and (JIA *et al.*, 2023).

Lenat and Guha published the general steps about the Cyc methodology (LENAT; GUHA, 1989). It is a process with three main phases: the first proposes a manual extraction of common sense knowledge through coding the explicit and implicit knowledge appearing in the knowledge sources without the help of natural language and learning systems (FERNÁNDEZ-LÓPEZ; GÓMEZ-PÉREZ, 2002); the second phase proposes a computer-aided extraction of common sense knowledge; and the last a computer managed extraction of common sense knowledge (TORRES *et al.*, 2024).

Grüninger and Fox proposed a methodology for ontology design and evaluation based on their experience developing the TOVE project (GRüNINGER; FOX, 1995). This methodology has six main steps: Motivating Scenario; Informal Competency Questions; First-Order Logic: Terminology; Formal Competency Questions; First-Order Logic: Axioms; and Completeness Theorems (TORRES *et al.*, 2024). Grüninger and Uschold also outlined a Formal Approach to Ontology Design and Evaluation (USCHOLD; GRUNINGER, 1996) called the seven-step method (JIA *et al.*, 2023).

Methontology is one of the most widely recognized methodologies. It offers a structured approach to building ontologies, encompassing specification, development, and maintenance (TORRES *et al.*, 2024). This methodology systematically includes stages such as knowledge acquisition, conceptualization, formalization, implementation, and ontology evaluation

(FERNÁNDEZ-LÓPEZ *et al.*, 1997). Lopez (LÓPEZ *et al.*, 2000) and Pérez (GÓMEZ-PÉREZ; ROJAS-AMAYA, 1999) also propose an extension to Methontology based on the concept of Ontological Reengineering.

Noy and McGuinness proposed Ontology Development 101, a simplified approach to guide novice ontology developers through the essential tasks for creating a fundamental ontology (TORRES *et al.*, 2024). It focuses on defining classes, properties, and instances, gradually formalizing the ontology (NOY, 2001). The On-To-Knowledge (OTK) methodology follows an iterative approach involving knowledge collection, modeling, implementation, and maintenance, emphasizing the reuse of existing ontologies and the use of automated tools (SURE *et al.*, 2004).

Lastly, by comparing these methods, (JIA *et al.*, 2023) proposes an improved seven-step method that leverages the strengths of the Methontology method and others to address the weaknesses of the original seven-step method.

In recent years, researchers have adopted deep learning and transformer-based models such as GPT and BERT to support ontology term extraction, relationship classification, and ontology extension. These models provide contextual understanding and facilitate new methodologies to integrate classical ontology engineering with generative AI capabilities. (CAPPELLI; SERUGENDO, 2025) explores LLM-based semi-automation using GPT-4o to accelerate ontology development while improving process efficiency and result quality. It uses diverse prompting to elicit structured LLM responses as foundations for ontology construction, which are subsequently enriched through manual curation.

Ontology construction processes have made several recent advances. Despite this, the literature emphasizes the role of domain experts and human supervision in ensuring semantic accuracy, contextual relevance, and interpretability. In addition, the new AI-based methods present some limitations, such as handling domain-specific terminology, mitigating model biases, or generating semantically coherent relationships.

The literature also emphasizes the importance of incorporating ontologies' evolution, reuse, and collaborative practices to maintain relevance and interoperability in dynamic, multi-domain applications. Challenges persist in managing semantic consistency during ontology updates. Domain-specific adaptations often struggle to balance automation and the need for expert contributions, limiting generalization.

There are no standardized, universally accepted criteria and metrics for comparing ontology development methodologies, making direct comparisons and the selection of ideal approaches difficult. However, the current state of methodology development reflects significant progress driven by AI integration but also reveals essential gaps in standardization, semantic accuracy, collaborative tools, data governance, and quality assurance.

2.2.1.4 Processes to Build Knowledge Graphs

Constructing KGs also involves several steps and viewpoints. To comprehend the overall process, (TAMAŠAUSKAITĖ; GROTH, 2023) conducted a systematic literature review to identify, describe, and integrate the key steps involved in creating a General Knowledge Graph. Additionally, (KEJRIWAL, 2019) investigated the steps involved in domain-specific knowledge graph construction, which is especially important because constructing knowledge graphs within organizations often poses domain-specific challenges.

(TAMAŠAUSKAITĖ; GROTH, 2023) followed a rigorous data collection process, selecting 57 relevant articles focusing on knowledge graph development. Data analysis of these articles, involving identifying and consolidating process steps, led to the formulation of a general knowledge graph development process. Based on this literature review, six key steps have been identified in the knowledge graph development. The process flow proposed by (TAMAŠAUSKAITĖ; GROTH, 2023) is presented in Figure 2.4:

- Identify Data (1): This step aims to identify a domain of interest, a data source, and a data acquisition method.

- Construct the Knowledge Graph Ontology (2): It builds the knowledge graph ontology, providing a top-level structure for the KG.

- Extract Knowledge (3): The main objective is to extract entities, relations between them, and attributes.

- Process Knowledge (4): This step's main objective is to ensure that unprocessed extracted entities, relations, and attributes are not ambiguous, redundant, or incomplete.

- Construct the Knowledge Graph (5): This step's final objective is to populate the KG and to ensure that the knowledge graph is accessible and available for use.
- Maintain the Knowledge Graph (6): KGs must be maintained over time to ensure accuracy and relevance. Thus, it is necessary to continuously monitor the knowledge graph, its usage, and relevant domain data sources, and update the knowledge graph as needed.

Finally, the authors make some essential reflections on the limitations of the process. They begin by guiding how to use the KG for the initial development and warn of the need for additional adaptations to apply existing knowledge graphs. They also explain that the process is a general guide and that the development team should conduct further research to construct it, such as the types of algorithms to be used or even the type of graph storage structure. There is also an observation regarding the direction of the research focus in the literature analysis, which requires a study of the results of applying this process in practice in organizations.

Academic processes for building Knowledge Graphs are often designed for specific, well-defined research projects. They lack formalism for the realities of a large government agency: continuous data updates, evolving business requirements, strict data governance, and constant collaboration between technical teams and non-technical legal experts. They are often presented as batch-like processes, unsuitable for incremental updates without complete recomputation.

These presented methodologies have specific gaps that need to be highlighted: there is a lack of business integration, the absence of this formalization is a critical flaw for real-world application, or the governance vacuum - how do these academic models handle data lineage, versioning, and compliance with legislation such as the General Data Protection Law (LGPD)? At the end, treating ontology construction as a mere subtask is problematic. In complex domains such as legal-investigative, ontology engineering is a parallel and iterative process of knowledge modeling, not a linear step in a data pipeline.

2.2.2 Named-Entity Recognition

Named-Entity Recognition is a crucial subtask of Natural Language Processing (NLP) that involves identifying and classifying proper names in text into predefined categories (LIU; WANG, 2017). NER employs a range of methodologies, from rule-based systems to deep lear-

ning models. Early NER systems relied on hand-crafted rules and dictionaries (JURAFSKY; MARTIN, 2009) and used pattern-matching techniques to identify entities based on these predefined rules, such as regular expressions and lexical patterns. While rule-based approaches are interpretable and straightforward, they are not scalable and can be challenging to maintain. They are limited by the rules' coverage, which means they often miss entities that do not fit predefined patterns (NADEAU; SEKINE, 2007).

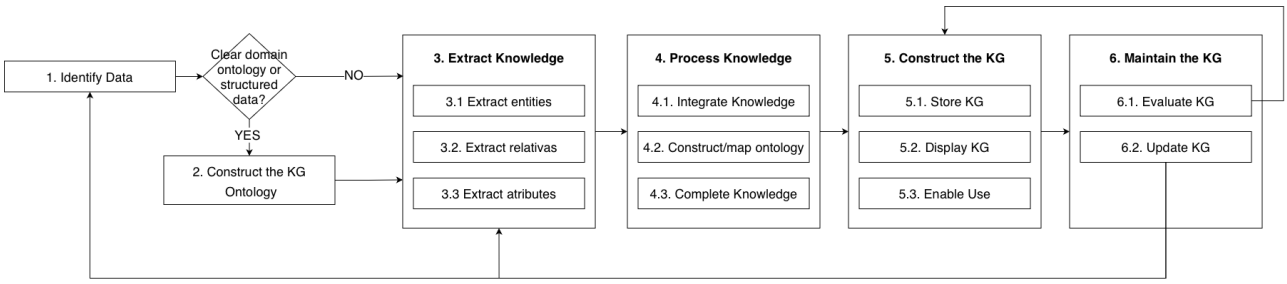

Figure 2.4: KG Development Process Proposed by (TAMAŠAUSKAITĖ; GROTH, 2023)

As computational power and data availability increased, machine-learning techniques became popular for NER. These approaches involve training statistical models on labeled datasets to identify entities. Some examples in this category are the Hidden Markov Models (HMM) (JURAFSKY; MARTIN, 2009), Conditional Random Fields (CRF) (SONG *et al.*, 2019), Support Vector Machines (SVM) (EKBAL; BANDYOPADHYAY, 2010), and a mix of models (TANG *et al.*, 2013). Machine learning models can generalize better than rule-based systems and handle a broader range of entity types. However, they require large annotated datasets and can be complex to train and tune.

Recently, deep learning models, such as Recurrent Neural Networks (RNNs) (CORBETT; BOYLE, 2018) (LIU *et al.*, 2017), Long Short-Term Memory Networks (LSTM) (HAMMERTON, 2003) (LYU *et al.*, 2017). Transformer-based models (e.g., Bidirectional Encoder Representations from Transformers - BERT) have led to significant improvements in NER. These models can learn complex patterns and dependencies in the text, making them effective for NER tasks (DEVLIN *et al.*, 2018).

Deep learning approaches have two main drawbacks: their need for large amounts of data and computational resources, and their tendency to be challenging to interpret due to their complexity. Combining rule-based and machine-learning methods can leverage the strengths of

both approaches. Hybrid systems might use rules to pre-process text and reduce noise, while machine learning models handle more complex entity recognition tasks (VEERASEKHAR-REDDY *et al.*, 2023). For instance, a hybrid approach could apply simple rules to identify common entity patterns and then use a machine learning model to refine the classification.

2.2.3 Match Entities and Record Linkage

Consolidating a unique entity dispersed across several databases is challenging. This problem occurs when these entities lack a unique identifier or when their data attributes are incomplete, outdated, or inconsistent. Dealing with different data formats and structures is often another challenge to be overcome to ensure data quality and consistency (CHRISTEN, 2012).

Several other challenges relate to privacy and confidentiality in data processing and to the computational complexity of data-matching algorithms. The computational complexity of data matching, for example, increases quadratically with the size of the databases, as each record from one database must potentially be compared with all records in the other (CHRISTEN, 2008). Personal information involved in matching processes also raises privacy and confidentiality concerns.

Dealing with the distributed storage of related information involves data integration. The first step of this process, often called schema matching (RAHM *et al.*, 2000), consists of identifying attributes and conceptual structures, such as ontologies, across multiple databases that store data from the same entity. Data matching is the second stage of the broader data integration process. It consists mainly of identifying and matching individual records from different databases that refer to the same entities (CHRISTEN, 2012). The third task combines pairs or groups of records recognized as matches into a single record. This step is called data fusion (BLEIHOLDER; NAUMANN, 2009).

Based on Newcombe's ideas (NEWCOMBE; KENNEDY, 1962), Fellegi and Sunter published a reference paper on probabilistic record linkage (FELLEGI; SUNTER, 1969). This work has served as a basis for many data-matching systems and software products. Their theory showed that an optimal probabilistic decision rule can be found under the assumption that

the attributes used to compare records are independent. They also proposed a mathematical model to provide a theoretical framework for solving the entity-matching problem.

The main objective of this process is to classify pairs of records in the product space S1 × S2 from two sources, S1 and S2, into two categories: MATCH and NON-MATCH. (FELLEGI; SUNTER, 1969) proposes a formula to consider ratios of probabilities 2.1, considering γ an arbitrary agreement pattern in a comparison space Γ, such as patterns representing simple or no agreement on the street name, and street number of a complete address. The given formula represents a ratio R of conditional probabilities (WINKLER, 2014).

$$R = P(\gamma \in \Gamma \mid M)/P(\gamma \in \Gamma \mid U) \tag{2.1}$$

Where:

- M means that comparing two records resulted in a match (i.e., they refer to the same entity).
- U means that the comparison resulted in a non-match (i.e., they refer to different entities).
- $P(\gamma \in \Gamma \mid M)$ is the probability that a given agreement pattern γ is observed, given that the records are a match.
- $P(\gamma \in \Gamma \mid U)$ is the probability that the same agreement pattern is observed, given that the records are not a match.
- R is the likelihood ratio that quantifies how much more likely the observed agreement pattern γ is under the assumption of a match (M) than under the assumption of a non-match (U).

Also, according to (WINKLER, 2014), the following rules give the decision. The thresholds defined by $T\gamma$ and $T\mu$ are determined by a priori error bounds on false matches and mismatches.

- If $R > T_\mu$, then designate the pair as a match
- If $T_\gamma \leq R \leq T_\mu$, then designate the pair as a possible match to be reviewed
- If $R < T_\gamma$, then designate a pair as a nonmatch

Another approach to record linkage uses machine learning, typically supervised or semi-supervised. Machine learning methods, particularly supervised and semi-supervised techniques, enable more nuanced and flexible entity matching than purely probabilistic frameworks. Machine Learning (ML) approaches typically involve crafting features that capture similarity metrics (e.g., string distances) and feed them into a learning model such as a decision tree, random forest, or gradient-boosting machine.

In supervised entity matching, a labeled dataset of confirmed matches and non-matches is used to train the model. The algorithm learns patterns in the feature space to predict the likelihood of a match for new or unseen pairs of records. Semisupervised techniques use smaller labeled datasets supplemented by large quantities of unlabeled data. Semisupervised models can improve performance even when manually curated labeled data is limited by propagating match and non-match decisions from labeled examples to unlabeled pairs.

While ML-based entity matching can achieve high accuracy, maintaining interpretability can be challenging. Feature importances and rule-based explanations often mitigate this issue, but the complexity of specific models may obscure how final match decisions are reached. Additionally, the transferability of the learned models to new domains may require domain-specific adaptation of feature engineering and parameter tuning to account for different naming conventions or attribute structures.

2.2.4 Emerging Paradigms with Large Language Models (LLMs)

This section covers using LLMs on Entity Resolution and Knowledge Graph Construction. Using ER to extract graph elements generally works as follows: a document or text excerpt is provided to the LLM with a prompt specifying the desired schema (the entities and relationship types to be extracted). The model then reads the text and generates a structured output, often in JSON format, listing the entities and relationships found. Related work shows that fine-tuning performance for this task is associated with the use of specific prompts (ALGOABRA, 2024).

There is also a specific line of research for building KGs from unstructured text. In this sense, LLMs can extract triples (subject-predicate-object) from legal documents, investigation

reports, or news articles to automatically populate and enrich the Knowledge Graph. Some researchers analyze how the current advances in foundational LLMs can be compared with specialized pretrained models, like Relation Extraction By End-to-end Language generation (REBEL) (CABOT; NAVIGLI, 2021), for joint entity and relation extraction, and to build pipelines for the automatic creation of Knowledge Graphs from raw texts (TRAJANOSKA *et al.*, 2023).

Despite their transformative aspect, the application of LLMs in a high-consequence context poses urgent challenges that must be carefully managed, such as:

- LLMs are prone to hallucinations, generating plausible information that is factually incorrect or unfaithful to the source text. A lack of factual reliability is a severe limitation in a domain where precision is imperative, and results can be used as evidence.
- The most capable models are generally proprietary and accessed via paid APIs, with costs that can become prohibitive for a public agency to process big data. The alternative, self-hosting open-source models, demands significant investment in hardware infrastructure and technical expertise for maintenance and optimization, resulting in a high cost.
- Data privacy and sovereignty are other vital challenges for public-sector applications. Submitting sensitive investigation data to third-party APIs in foreign jurisdictions raises serious concerns about privacy, security, and national sovereignty.

2.2.5 Network Analysis and Anomaly Detection with Graph Neural Networks (GNNs)

Fraud data modeling often results in heterogeneous graphs with multiple node and edge types, and GNN architectures have evolved to address this complexity. The most essential GNN architectures include Graph Convolutional Networks (GCN) (ZHANG *et al.*, 2019), Graph Sample and aggregate (GraphSAGE) (ZHANG *et al.*, 2022), Graph Attention Networks (GAT) (VRAHATIS *et al.*, 2024), and Heterogeneous Graph Transformer (HGT) (HU *et al.*, 2020).

Graph Convolutional Networks operate transductively, meaning they learn an embedding for each specific node they encounter during training (ZHANG *et al.*, 2019). This characteristic makes it suitable for classifying nodes within the graph. Imagine, for example, a graph with

1 million users, 10,000 of whom are labeled "fraudulent."After training, the model will have generated embeddings for all 1 million users and will be able to use them to predict the labels of the remaining 990,000 users who lacked labels. It is important to note that when users 1,000,001 register tomorrow, the trained GCN does not have an embedding for them. The model does not know how to classify them. Adding this node to the graph and retraining the entire model from scratch would be necessary to include them.

Furthermore, nodes within the same community (e.g., a fraud ring) will have very similar embeddings, so it is possible to use a clustering algorithm (e.g., k-Means) on these embeddings to identify and isolate these groups. GCN also works in link prediction, estimating the probability of an edge between two nodes that already exist in the graph but are currently unconnected (ZHANG *et al.*, 2019).

GraphSAGE is a graph neural network architecture that improves on the GCN idea. It has an inductive nature, i.e., instead of learning a specific embedding for each node in a fixed graph, GraphSAGE learns a general aggregation function that knows how to collect information from a node's neighborhood and generate an embedding for it, even if the node has never been seen during training (ZHANG *et al.*, 2022).

Instead of using all neighbors as in GCN, GraphSAGE samples and randomly selects a fixed number of neighbors for each node, avoiding the need to consider the entire neighborhood, making the process computationally feasible and scalable. The information from the sampled neighbors is then aggregated using an aggregation function, such as averaging the neighbor vectors (similar to GCN (ZHANG *et al.*, 2022)). The node's embedding is then updated by combining its current embedding with the aggregated vector of its neighborhood.

GraphSAGE is often used in Real-Time Fraud Detection. It can, for example, generate an embedding for a new transaction or user in near real-time and classify it as potentially fraudulent based on its connection to the existing network, providing real-time detection without retraining the model.

The central idea of GAT is simple: not all of a node's neighbors are equally important. While GCN treats all neighbors democratically (calculating a simple average) and GraphSAGE aggregates neighbors uniformly (whether by average, maximum, etc.), GAT introduces the attention mechanism, a concept borrowed from Natural Language Processing. This mechanism

allows the model to learn, for each node, to assign different weights to each neighbor during the aggregation process (VRAHATIS *et al.*, 2024).

Updating a single node begins with a linear transformation, where the features of all nodes undergo a linear transformation (they are multiplied by a learned weight matrix W). It projects the features into a higher-dimensional space where patterns can be more easily found. After that, the attention coefficients ((α_{ij})) are calculated for each neighbor j. This score represents the importance of node j's features to node i. This calculation is performed by a small, single-layer neural network trained alongside the rest of the model (VRAHATIS *et al.*, 2024).

Raw attention scores are difficult to interpret. To make them comparable, the model applies the softmax function to all scores of node i and its neighbors. It transforms the scores into a probability distribution where the new attention coefficients (α_{ij}) sum to 1. Instead of a simple average, GAT calculates a weighted average of the transformed features of the neighbors. The weight of each neighbor is precisely the attention coefficient (α_{ij}) calculated in the previous step. The result of this weighted sum becomes the new representation (the new embedding) for node i. Typically, a nonlinear activation function (such as LeakyReLU) is applied at the end.

GAT focuses on the relationships that truly matter in defining fraudulent behavior, ignoring the hundreds of legitimate interactions that could confuse other models. Another positive point is Interpretability (Explainability), as it allows visualization of attention weights and answers questions such as "Why did the model pay 80% of its attention to connecting to a particular device, which is associated with a known fraud ring?"

Heterogeneous graphs present two significant challenges. The first is the varying node types, each with a different feature space and features. For example, those describing a Person are entirely different from those describing a Company. They cannot simply be mixed. The second is the other edge types, each representing a different interaction with different importance.

HGT addresses these problems by adapting the Transformer architecture to the domain of heterogeneous graphs (HU *et al.*, 2020). HGT also uses multiple attention mechanisms, one for each possible interaction type in the graph. The central idea is the meta-relation, a tuple that describes a path type in the graph: <Source Node Type, Edge Type, Target Node Type>. HGT learns an independent and specialized set of attention weights for each meta-relationship.

The process of updating a node begins with identifying the meta-relationship. For each neighbor, the model first identifies the meta-relationship connecting them. Then, instead of using a generic attention formula, HGT uses explicitly created weight matrices (W_query, W_key, W_value) for the meta-relationship (HU *et al.*, 2020). The neighbor's message is calculated using this specialized attention mechanism, and the messages received from all neighbors are grouped by meta-relationship type and aggregated (usually by an average). Finally, the target node's embedding is updated by combining the aggregated messages from each meta-relationship type.

Heterogeneous models with attention and multiple aggregators are frequently used in the literature to capture the semantics of distinct relationships effectively (WU *et al.*, 2024)(JIANG *et al.*, 2021). Recent literature also presents specialized architectures for fraud detection in specific domains:

- MAFI (Multiple Aggregators and Feature Interactions) (JIANG *et al.*, 2021): Focused on e-commerce, it uses multiple aggregators specialized by relationship type, attention mechanisms at the aggregator and relationship level, vectorial feature interactions, and a trainable sampler to filter out cloaking.

- LIFE (Live-streaming Fraud Detection) (LI *et al.*, 2021): A heterogeneous model with attention, developed for the Taobao platform. It employs semi-supervised label propagation to handle label sparsity and relationship-aware aggregation to capture specific semantics.

The training strategy of GNN models depends on the availability of fraud labels, which are typically scarce and expensive. Most research formulates fraud detection as a supervised binary classification, where nodes or transactions are classified as fraudulent or legitimate (JIANG *et al.*, 2021). Semi-supervised approaches are frequently employed to mitigate the dependence on large volumes of labeled data, such as Label Propagation, which uses the graph structure to propagate known fraud signals to unlabeled neighboring nodes. The LIFE model (LI *et al.*, 2021) is an example of an application in heterogeneous graphs. Alternatively, Multi-Task Learning, which jointly trains the main task (fraud detection) with auxiliary tasks (e.g., feature prediction, edge reconstruction), improves generalization. Despite these advances, the practical application of GNNs for fraud detection faces significant challenges, frequently identified in the

literature (ZHOU *et al.*, 2024) (LI *et al.*, 2021) (XIE *et al.*, 2024):

- Class Imbalance: Fraud typically represents less than 1% of transactions (ZHOU *et al.*, 2024), with ratios of 1:100 to 1:1000 being common.
- Scalability for Large Graphs: The need to process graphs with millions or billions of nodes/edges and perform real-time inference imposes severe memory and computational constraints.
- Hidden Fraud: The presence of unlabeled fraud ("label noise") in the training data contaminates learning legitimate patterns, especially critical in unsupervised methods.
- Noise and Information Overload: Knowledge graphs may contain irrelevant nodes, spurious relationships, or noisy attributes (SHAO *et al.*, 2025).
- Explainability and Adversarial Attacks: In regulated environments, justifying model decisions is an operational and compliance requirement (XIE *et al.*, 2024). The literature still lacks evidence of robust defenses against targeted adversarial attacks (e.g., manipulation of graph structure or attributes) in GNNs for fraud detection, with more focus on robustness to noise.

In a legal context, each inferred link must be auditable and defensible. GNNs are challenging to interpret, operating as black boxes. Their opacity is a critical flaw for this application, which requires transparency. Furthermore, there are problems with cold Start and Labeled Data, as GNNs require vast amounts of labeled fraud data for supervised training. The cold start problem is a classic and fundamental challenge in machine learning systems. It occurs when a system fails to make valid inferences or predictions for users or items about which it has not yet collected sufficient data. The term is an analogy to a car engine that struggles to start on a freezing day, but runs perfectly once it reaches the optimal operating temperature. In our context, the engine is the AI model, and the optimal temperature is a sufficient volume of historical data.

Labeled fraud data is scarce in government, and high-quality labels are difficult to obtain. An initial rule-based approach, grounded in ontology and expert knowledge, circumvents this limitation at a time when government maturity is still low. New heuristics based on GNN are

expected to be implemented after developing this basic expertise, since they are even included in the methodology for constructing knowledge graphs.

2.3 ASSET CONCEALMENT INVESTIGATION TOOLS

The following subsections will present a non-exhaustive list of tools commonly used in asset concealment investigations. The set includes data analysis and link-detection solutions, OSINT tools, Knowledge Graph platforms, and systems for analyzing financial transactions.

Each tool category plays a specific and complementary role in the asset investigation and recovery ecosystem. Together, these tools provide data collection and integration, identification of complex relationships and patterns, analysis of textual and financial evidence, and consolidation of entities to avoid information dispersion or duplication.

The coordinated adoption of these categories can improve investigations, reduce analysis time, increase the accuracy of asset identification, and, consequently, optimize the outcomes of asset recovery actions. However, integrating these tools also creates a gap that needs to be addressed.

2.3.1 Data Collection and Integration Tools

Dispersed, heterogeneous databases pose one of the biggest challenges in asset concealment investigations. Data collection and integration tools are used to unify information from bank records, notary offices, government agencies, and online platforms, among other sources. In addition, they standardize data formats, reducing noise caused by different storage models, terminologies, and record quality levels. They also enable the analysis of large volumes of data more efficiently, offering Extract, Transform, Load (ETL) or Extract, Load, Transform (ELT) mechanisms.

Without a collection and integration layer, any investigative procedure would be fragmented, subject to inconsistencies, and loss of information essential to locating misappropriated or undervalued assets. Several tools in this category exist, such as Apache NiFi (APACHE SOFTWARE FOUNDATION, 2025) or Qlik Data Fabric (QLIK, 2025). They are essential for

ingesting and orchestrating data from different sources (banks, notaries, government agencies) into a standardized format, facilitating the detection of disparities or potential concealment.

Apache NiFi is a dataflow system. It is based on concepts of flow-based programming and supports scalable, directed graphs for data routing, transformation, and system mediation logic (APACHE NIFI TEAM, 2025). It operates asynchronously, and the steps can be executed in parallel, as they are independent. More than 250 pre-defined tasks are available to users, and it is also possible to define new modules for non-standard tasks (WNęK; BORYłO, 2023).

Qlik Data Fabric is a machine-enabled data integration architecture that utilizes metadata assets to unify, integrate, and govern disparate data environments. By standardizing, connecting, and automating data management practices and processes, data fabrics improve data security and accessibility and provide end-to-end integration of data pipelines and on-premises, cloud, hybrid multi-cloud, and edge device platforms (QLIK, 2025).

A Data Fabric facilitates a distributed data environment where data can be ingested, transformed, managed, stored, and accessed across various repositories and use cases, such as BI tools and operational applications. It achieves this by employing continuous analytics over current and inference metadata assets to create a web-like layer that integrates data processes and the many sources, types, and locations of data. It also employs modern methods such as active metadata management, semantic knowledge graphs, embedded machine learning, and AutoML (QLIK, 2025).

Although tools like Apache NiFi are useful for orchestrating data flows and executing ETL/ELT processes, their function is primarily "data plumbing". They are agnostic to the content they transport, moving bytes from one system to another with high efficiency, but without a semantic understanding of the data. These tools do not natively possess a domain model (such as an ontology) to unify concepts, resolve ambiguities, or infer new relationships. In short, they move and transform data but do not create knowledge. This limitation highlights the need for a higher layer of intelligence, such as the one proposed in this ecosystem, which uses a domain ontology to give meaning to and connect the data that these tools integrate.

2.3.2 Open Source Intelligence Tools - OSINT

OSINT tools and methods enable the collection of public information from the Internet, including social networks, government websites, forums, and open data databases. This category of solutions is important because it complements official (and potentially outdated) data, offering updated insights into lifestyle, family ties, undeclared transactions, assets (vehicles, properties, etc.), and business connections. This way, OSINT tools can collect public information online, such as social media, domain registries, and the dark web. They may reveal evidence of misreported assets or connections between individuals and companies not included in official records. They also enable the identification of suspicious behaviors, such as expenses inconsistent with declared income.

There are tools such as Maigret (SOXOJ, 2025), which collect a dossier on a person by username. It checks for accounts on more than 2,500 sites and gathers all the available information from web pages. Another essential tool is the CrimeWall by Social Links (SOCIAL LINKS, 2025). This solution streamlines the investigation process and combines in-depth open-source data extraction with user-friendly visualization and analysis features. Finally, we can mention Social Searcher, an online tool for searching for public posts from users on social networks such as Twitter, Facebook, and Instagram (SOCIAL SEARCHER, 2025).

These tools facilitate open data collection and analysis; however, they do not include the domain knowledge needed for Asset Concealment Investigations, which require specialized expertise across several areas. While OSINT tools can help identify links between individuals, companies, and assets, they cannot interpret complex financial schemes or follow the assets. Effective asset tracing demands data collection, contextual understanding, and investigative methodologies.

Open Source Intelligence tools are valuable for collecting open data from social media or the web. However, they often operate as raw data collectors, disconnected from structured and confidential government databases. They lack a framework for automatically merging the collected public data and a public agency's internal records. An investigator may find a name on a social network, but connecting that digital persona to a CPF (Brazilian individual taxpayer registration number) or a company registration in an internal database remains a manual,

error-prone process. The ecosystem proposed in this thesis addresses this gap by providing a unified platform for integrating and analyzing both internal and external data within a single knowledge graph.

2.3.3 Semantic Analysis and NLP Tools

In asset concealment investigations, many clues can be found in unstructured text documents, such as testimonies, extracts, contracts, social media posts, and news reports. NLP and semantic analysis methods extract named entities (people, companies, locations, assets) in an automated manner, which is essential for scaling up investigations of large amounts of text. Additionally, they resolve ambiguities such as different spellings of names, the use of acronyms and specific terms, and the linking of textual references to known entities in structured databases.

Understand the context through semantic analysis algorithms that relate text to possible evidence of ownership, possession, or the ultimate beneficiaries of assets. In this way, NLP introduces a linguistic component to the investigation, which is vital for correlating reports and documentation and for identifying assets spread across multiple textual records.

Several libraries exist for Natural Language Processing, named entity extraction (people, places, organizations), and semantic analysis of texts, such as spaCy (SPACY, 2025) and NLTK (BIRD *et al.*, 2009). Typically, these libraries provide algorithms and models that must be adapted to a specific case or certain data sets under analysis. They can assist in unifying names and addresses, detecting aliases, and identifying mentions of assets in documents or social networks.

SpaCy can identify different types of named entities within a document by leveraging its model to generate predictions. However, these models rely on statistical methods and are heavily influenced by the training data on which they were built; their accuracy is not always flawless. As a result, it may be necessary to fine-tune the model to improve performance. Additionally, version 3.0 introduces transformer-based pipelines, bringing spaCy's accuracy to the current state of the art (SPACY, 2025). The named entity recognition accuracy of spaCy on the OntoNotes (WEISCHEDEL *et al.*, 2011) is presented in Table 2.2 (SPACY, 2025).

Libraries like spaCy are powerful toolkits for NLP, allowing the extraction of named entities

Table 2.2: Named entity recognition accuracy of spaCy on the OntoNotes 5.0 (SPACY, 2025).

NER System	**OntoNotes** (WEISCHEDEL *et al.*, 2011)
spaCy RoBERTa (2020)	89.8
Stanza (StanfordNLP) (QI *et al.*, 2020)	88.8
Flair (AKBIK *et al.*, 2018)	89.7

(NER) from unstructured texts. However, they are low-level components that require significant custom development to be effective in domains as specific as legal proceedings or financial reporting. They identify a "name,"but do not resolve homonym ambiguity or connect the extracted entity to a master record. They are one piece of the puzzle, not the complete solution. The absence of an entity resolution layer and a knowledge graph to contextualize the extracted information makes their isolated use insufficient for large-scale investigation, justifying the need for the AI models and integrated architecture proposed in Chapter 7.

2.3.4 Entity Resolution and Record Linkage Tools

Techniques such as Entity Resolution (ER) and Record Linkage (RL) address the problem of matching records that contain information about the same entity across different databases but appear with various spellings, accents, surnames, or identifications. These methods build a single view for each individual or company. They facilitate the centralized analysis and the detection of hidden assets linked to name variations.

There are several tools for ER and RL. Senzing, for example, is a platform specialized in resolving identities between large databases. It uses advanced heuristics and machine learning to handle variations in spelling, typos, surname changes, etc. Splink is a Record linkage library for Python that uses probabilistic techniques and machine learning algorithms to cross-reference data across multiple sources. It can be customized to include specific weights in fields such as name or address (LINACRE *et al.*, 2022).

Despite the availability of tools like Senzing and Splink, effective Entity Resolution and Record Linkage require a well-defined strategy tailored to the data sources and domain-specific challenges. The success of these techniques depends not only on the algorithms and heuristics, but also on the context, the quality of the input data, and the ability to handle inconsistencies across different systems.

High-performance platforms for Entity Resolution, such as Senzing, offer high performance for matching records in large volumes of data. However, many operate as "black boxes,"using proprietary heuristics and algorithms. This lack of transparency is an essential disadvantage in the investigative and legal context, where every step of the analysis and every inferred connection must be auditable and explainable. Furthermore, they are point solutions that need to be integrated into a broader analytical workflow. The ecosystem proposed in this thesis offers an alternative by developing customized, transparent AI models natively integrated into the domain ontology, ensuring that the logic behind the unification of each entity is transparent and defensible.

2.3.5 Knowledge Graph and Ontology Platforms

Using knowledge graphs and ontologies brings a semantic approach to data processing. Instead of storing information relationally, a knowledge graph aggregates structures and relationships that can be inferred at analysis time, taking into account concepts and hierarchies. It represents complex relationships more naturally, allowing the modeling of links such as "person A is the owner of company B"and "company B has a bank account in country X. "It also facilitates the inference of new facts from semantic rules or domain-specific ontologies and improves data consistency because it allows the creation of controlled vocabularies and ontological schemas, which reduce ambiguities that hinder research.

These platforms provide data enrichment and deeper analysis capabilities, allowing investigators to find correlates of assets that would otherwise go unnoticed in traditional database systems. In this category, we can cite tools for building ontologies, such as Protege, and graph databases for storing knowledge graphs, such as Neo4j and Amazon Neptune.

Protégé is an open-source platform with tools for building domain models and knowledge-based applications using ontologies (STANFORD UNIERSITY, 2025). It fully supports OWL 2 and direct in-memory connections to description logic reasoners such as HermiT and Pellet.

Neo4j is a graph database management system designed to store, manage, and analyze highly connected data efficiently (NEO4J Inc., 2025). Unlike traditional relational databases, which rely on tables and joins, Neo4j leverages a native graph model, representing data as

nodes (entities) and relationships (connections). It enables fast, flexible querying of complex relationships, making it useful for applications such as fraud detection using knowledge graphs. Neo4j helps investigators visualize and traverse these complex networks, designed to obscure the true ownership of assets, by mapping relationships among individuals, companies, bank accounts, and transactions.

Tools like Neo4j and Protégé are fundamental technologies, but not the solution. Neo4j is the "engine,"but it needs integrated data and a "map"(the ontology) to be useful. Protégé is an ontology editor. The mere existence of these tools does not solve the central problem of asset concealment. The main contribution of this thesis is not the use of these technologies, but rather the construction of the entire end-to-end process: the methodology for creating the ontology (Chapter 4), the process for building the graph (Chapter 5), the specific domain ontology (Chapter 6), and the AI models for populating and analyzing the graph (Chapter 7). The proposed ecosystem is the complete solution, not just its isolated parts.

2.3.6 Link Analysis Tools

Asset concealment investigators often use these tools to uncover hidden connections between individuals, companies, and assets. Link Analysis tools allow the identification of complex relationship patterns, which can help detect links between people and assets that are not explicit in structured databases. They also map transaction routes and capital movements, helping to understand suspicious financial flows.

Governmental agencies broadly use tools such as i2 Analyst's Notebook (N. HARRIS COMPUTER CORPORATION, 2025) and Maltego Graph (MALTEGO TECHNOLOGIES, 2025) to analyze graph data. These tools, for example, allow visualization of networks of relationships involving financial transactions, properties, contracts, or business partnerships. i2 Analyst's Notebook uses Entity-Link-Property (ELP) methodology to find connections between data entities. It can analyze and visualize large volumes of data from multiple sources to help users identify, predict, and prevent criminal and fraudulent activities. Maltego Graph is one of the world's most used cyber investigation platforms (MALTEGO TECHNOLOGIES, 2025). It enables seamless integrations with data, social network mapping, person-of-interest investigations,

and cryptocurrency transaction tracking, among other functionalities.

Both tools are essential in the investigative process. However, they focus on the final stage, which involves analyzing and visualizing data that has already been processed and stored in an appropriate format. In this sense, without properly carrying out this preliminary phase, the results of using these tools are minimal.

Analysis of tools such as IBM i2 and Maltego reveals that their primary focus is on the visualization and manual exploration of already processed and integrated data. They are powerful interfaces that help an analyst visualize the graph and navigate known connections. However, they have limited capacity for proactively and automatically discovering complex, non-obvious patterns in large-scale networks. They are not, in essence, inference engines that suggest new hidden connections. It justifies the approach of this thesis, which develops AI models that go beyond visualization by actively searching for implicit relationships in the graph.

2.3.7 Financial Transaction Analysis and Anti-Money Laundering Tools

Financial transaction analysis, transaction monitoring, and AML solutions are central to identifying money laundering and illicit financial flows. These tools detect anomalies in transactions, such as exorbitant amounts, serial transfers, or the use of offshore accounts; analyze suspicious movement patterns; and associate bank accounts, credit cards, and digital wallets with risk profiles. Applying AML mechanisms in the context of asset recovery helps to trace the path of money, illuminating routes of diversion or conversion of assets into forms of value that are difficult to detect.

In this way, we can cite Quantexa, which is focused on entity resolution, fraud detection, and money laundering (QUANTEXA, 2025). It uses graphs, machine learning, and correlation rules to identify suspicious transactions and potential real business owners (Figure 2.5). This type of tool is the closest to the proposed ecosystem. However, it is a high-cost corporate tool with an architecture not publicly available. A more detailed analysis of its operation was impossible due to the difficulty of obtaining additional information. Based on the public information available, it is impossible to infer how the process of identifying data sources, acquiring and preparing data before visualization and analysis, the most critical step in the ecosystem proposed in this

work, is carried out.

Other Quantexa-like tools used in financial institutions to analyze patterns, identify atypical transactions, and support asset tracing and regulatory compliance investigations face the same proprietary access problem. Some examples are Oracle Financial Crime and Compliance Management (FCCM) Solutions (ORACLE, 2024), and NetReveal's BAE Systems (BAE SYSTEMS, 2024).

Tools in this category align closely with the proposed ecosystem. It is important to note, however, that these tools are proprietary, meaning their methods, architecture, algorithms, and models are not publicly available. There are no further details on how the tool is given context or on the use of ontologies or KGs. This reality even prevents the comparison of the proposed ecosystem and the models that have been developed. In addition, the high cost of some of these tools may discourage many governments from acquiring them.

Figure 2.5: Quantexa User Interface (QUANTEXA, 2025)

Commercial AML solutions like Quantexa represent the closest approximation to the ecosystem proposed here. However, their limitations for the public sector motivate this thesis. Their proprietary nature and closed architecture impose high cost and flexibility barriers, but, more critically, they clash with the need for transparency and auditability. The black box nature

of their algorithms is incompatible with the context, where every inference must be subject to scrutiny. Therefore, this thesis does not seek to replicate such tools, but rather to offer a strategic alternative that is better adapted, more transparent, and more sovereign to the reality of the Brazilian State.

2.4 TOOLS SUMMARY

As presented, ecosystems dedicated to investigating asset concealment fraud employ various tools, methods, and techniques. Table 2.3 summarizes the main categories of tools associated with this area of study, presented throughout this section, outlining their strengths and potential limitations.

2.5 SUMMARY

A comprehensive literature review in asset concealment investigation reveals a vital paradox. While the individual components needed to combat asset obfuscation, such as data integration techniques, graph databases, and advanced AI models, are well-researched and available, the government's ability to leverage them effectively remains limited.

This thesis argues that this failure stems not from a lack of tools or method, but from a profound fragmentation gap which is threefold: semantic fragmentation, where data from disparate sources lack an ordinary meaning, impeding interoperability; methodological fragmentation, where academic processes for knowledge engineering fail to meet enterprise-level requirements for governance, iteration, and multidisciplinary collaboration; and architectural fragmentation, where a collection of powerful but disconnected tools fails to constitute a coherent, intelligent ecosystem.

Semantic fragmentation is one of the most fundamental problems because integrating records from different government agencies becomes manual, error-prone, and unsustainable without a formal and shared vocabulary. The lack of a domain ontology for asset obfuscation forces each new investigative effort to 'reinvent the wheel,' leading to inconsistent data models and hindering inter-agency collaboration.

Table 2.3: Comparative Analysis of Tool Categories

Tool Category	**Examples**	**Strengths**	**Critical Limitations**
Data Integration	Apache NiFi, Qlik Data Fabric	High performance in orchestrating data flows (ETL/ELT).	Semantically agnostic; does not understand the content of the data, requiring a superior ontological layer.
Open Source Intelligence (OSINT)	CrimeWall, Maigret, Social Searcher	Valuable for collecting public data (social networks, open web).	Operate as raw data collectors, disconnected from governmental databases. The fusion of public and internal information remains a manual process.
Semantic Analysis & NLP Tools	spaCy, NLTK	High accuracy in Named Entity Recognition (NER) from unstructured texts.	They are low-level components that do not resolve ambiguities (homonyms) or link extracted entities to a master record. They require a higher layer of resolution and contextualization.
Entity Resolution	Senzing	High accuracy in large-scale record linkage.	Black-box nature; proprietary heuristics hinder the auditability and explainability required in a legal context.
Knowledge Graph & Ontology Platforms	Neo4j, Protégé	They are essential enabling technologies; the Neo4j for storing/-querying the graph and the Protégé for building the semantic map.	They do not constitute the business solution by themselves. The thesis's contribution lies in the end-to-end process that uses them, not in the isolated use of the tools.
Link Analysis	Maltego, i2 Analyst Notebook	Powerful interfaces for manual visualization and exploration of networks.	Limited capacity for proactive and automated discovery of complex patterns; they are exploration tools, not inference tools.
AML Platforms	Quantexa, FCCM, Net Reveal's	Integrated solution that combines entity resolution and network analysis.	Closed architecture, high cost, and lack of customization for the specifics of data sources and regulations of the Brazilian public sector.

Methodological fragmentation manifests as a disconnect between the Knowledge Graph construction processes proposed in academia and the operational realities of a government entity. Existing methodologies often neglect the need for an iterative life cycle, incorporating ongoing data governance and structured collaboration with non-technical domain experts, such as attorneys and legal analysts, and treat Knowledge Graph construction as a linear, technically focused project.

Finally, architectural fragmentation reflects the current state of the tools market. There are point solutions for each process step—ETL, graph storage, visualization, entity resolution—but a unifying design that integrates them into a cohesive, intelligent workflow is lacking. Commercial platforms that attempt to offer an end-to-end solution do so as proprietary black boxes. This approach is incompatible with the public sector's transparency, auditability, and technological sovereignty requirements. The literature, therefore, points to the critical need not for another isolated tool but for an ecosystem. This socio-technical framework combines an open technological architecture with robust methodological processes and a formal semantic basis.

This thesis aims to address three fundamental research gaps based on this critical synthesis:

- The Semantic Gap: Despite the need for a unified data model for complex financial investigations, the literature lacks a formal, standardized, and reusable semantic schema for the asset concealment domain. The research conducted for this thesis itself confirmed that no public ontologies exist in this area. This absence is a primary obstacle to large-scale data integration and interoperability across agencies and jurisdictions.

- The Methodological Gap: Existing academic methodologies for building KG are insufficient for a government agency's dynamic, high-stakes environment. The methods are often linear and technically focused, lacking built-in processes for continuous data governance, iterative ontology refinement, and structured collaboration with non-technical domain experts.

- The Architectural Gap: There is a significant disconnect between the most powerful AI techniques (GNNs, LLMs) and the practical requirements of the legal-investigative domain. The black-box nature, high computational costs, the need for massive labeled datasets, and data privacy concerns associated with these models make them unfit for direct application by a public agency like the AGU. The field lacks models that are simultaneously effective and meet the rigorous criteria of explainability, auditability, and legal defensibility.

To address these identified deficiencies, this thesis proposes a multifaceted contribution. In response to the Semantic Gap, Chapter 6 introduces the first formal and reusable ontology for the asset concealment domain. To close the Methodological Gap, Chapters 4 and 5 detail

innovative, data-driven, and enterprise-ready processes for building ontologies and Knowledge Graphs that incorporate governance and expert collaboration as central and ongoing activities. Finally, to bridge the Architectural Gap, Chapter 3 presents the Conceptual Framework and 7 presents the implementation of a computational ecosystem that utilizes customized, transparent AI models and an explainable rules-based inference framework, providing a robust, practical, and legally defensible solution for the Brazilian public sector.

CHAPTER 3

CONCEPTUAL FRAMEWORK

This chapter establishes the theoretical basis, architectural structure, and socio-technical logic that underpin the proposal of this thesis. The ecosystem outlined here serves as the critical strategic link between the fragmented government data phenomenon and the methodological solution based on Knowledge Graphs. It serves as a dynamic map that transforms isolated, inert administrative records into actionable legal intelligence. The strategic role of this chapter is to lay the formal groundwork for a paradigm shift in asset recovery: the transition from a reactive search model—limited by exact keyword matching and relational schema rigidity—to a proactive detection paradigm based on network anomaly detection, semantic inference, and graph topology analysis.

The investigation of asset concealment is a problem characterized by adversarial dynamics, where the subjects of investigation actively modify their behavior to evade detection. Consequently, the architectural response must be isomorphic to the problem: it must be relational (Graph-based), robust against obfuscation (Semantic), and adaptable (Schema-flexible). While Chapters 1 and 2 identified the semantic gaps (lack of shared vocabulary) and architectural voids (absence of integrated tools), this chapter proposes the systemic activities and structural couplings necessary to unify them.

The design of this framework directly addresses the critical gaps identified in the literature review (Chapter 2). The analysis revealed a threefold fragmentation in the state of the art:

- Semantic Fragmentation: The absence of a shared, computable vocabulary for integrating investigative data across agencies and borders, leading to data silos where terms like "ownership"or "control"have divergent meanings.

- Methodological Fragmentation: The lack of corporate processes that combine governance, iterative refinement, and collaboration with domain experts. Academic literature often treats graph construction as a linear, one-off experiment, neglecting the cyclical nature

of intelligence work.

- Architectural Fragmentation: The existence of isolated tools rather than a cohesive ecosystem that supports the full lifecycle of intelligence—from raw data ingestion to court-admissible evidence.

In response, the architecture proposed here directly reflects the Design Science Research principles of creating purposeful artifacts to solve relevant organizational problems. It structures the interaction between the social system (investigators, lawyers, domain experts) and the technical system (algorithms, databases, interfaces), acknowledging that technology alone cannot solve the problem of asset concealment without accounting for the law's interpretative context.

The following sections will detail each framework component, analytically justifying its design choices with respect to the research objectives. We will present the Semantic Foundation, the domain ontology that ensures interoperability and defines the system's worldview. Next, we will describe the Data-to-Knowledge Pipeline, the processing engine that transforms raw data into actionable intelligence, highlighting the necessity of a Composite AI to handle decision logic. We will then dissect the Knowledge Graph itself and compare its efficacy with relational models. Finally, we will address the Exploration and Application Layer, focusing on translating intelligence into court-admissible evidence, and the Governance and Continuous Improvement Cycle, which ensures value generation and ecosystem sustainability. By the end of this chapter, the reader will have a clear understanding of the reference architecture that underpins the practical implementation and validation presented in Chapter 7.

3.1 SEMANTIC FOUNDATION

The Asset Concealment Domain Ontology (ACDO) is a formal and explicit specification of the concepts, properties, and relationships (GRUBER, 1993) that define the domain of asset concealment investigation. It serves as a shared, computationally interpretable vocabulary that enables systems and analysts to understand the connections among all entities in the system unambiguously. This foundation allows for proper semantic integration, transforming a set of heterogeneous, disparate data repositories into a cohesive, navigable knowledge network.

ACDO is the central semantic pillar in this proposal. This approach addresses the semantic fragmentation gap defined in Chapter 2, which identified the absence of a unified data model for asset concealment investigations. In the context of government data integration, silos are often not just technological but conceptual; different agencies define "ownership,partnership,"or "asset"differently. For example, a tax agency may define an asset based on its declared value, while a property registry may define it based on the deed title. Without a unifying ontology, merging these datasets results in semantic dissonance.

The ontology acts as the semantic interlingua to ensure interoperability and conceptual consistency across the ecosystem. Without ACDO, integrating data from different government sources becomes a manual, error-prone process of "join-and-hope,"dependent on ad hoc interpretations that result in inconsistent data models and hinder collaboration across research fronts. Furthermore, the choice of an ontology-based approach is a strategic response to the limitations of traditional relational database schemas. In a relational model, the "schema"is rigid and table-centric, optimized for storage efficiency rather than relationship discovery. When investigating fraud, the crime's schema is often unknown a priori. An ontology provides the necessary abstraction layer, allowing the system to model complex, real-world interactions—such as a "straw man"acting as a proxy for a beneficial owner—without being constrained by the rigid foreign-key structures of SQL databases.

It is imperative to address the limitations and conceptual boundaries of this approach. As highlighted in recent sociotechnical scholarship, data integration is not a politically neutral act; it involves ontological politics. By defining the ontology, we are actively constructing a specific "vision"of the world—deciding which relationships matter (e.g., "family ties,business partnerships") and which do not. This design choice carries inherent biases. For instance, defining a "risk"indicator based on specific family structures or geographic locations enacts a particular worldview that guides the investigation.

The performativity of the ontology means that the system does not just describe the criminal world; it helps construct the lens through which law enforcement perceives it. If the ontology heavily weights family connections as a fraud vector, investigators will be algorithmically nudged to scrutinize family members, potentially overlooking other non-familial criminal networks. This phenomenon demonstrates that an ontology is an assemblage of interests translated into code.

To mitigate the risks of "black-box"bias, the ACDO design follows a transparent, white-box methodology. The construction of this semantic artifact, developed in Chapter 6, followed a data-driven methodological process detailed in Chapter 4. The adopted methodology was based on analyzing real data clusters and capturing tacit knowledge from AGU domain experts. It ensures that the ontology reflects the legal and investigative reality rather than arbitrary technical assumptions. Furthermore, the process included an alignment step to ensure interoperability with established reference ontologies in the financial and social relations domains.

It is essential to note that, in this framework, the ontology proactively guides two of the ecosystem's main mechanisms. It defines the target structure for the Data-to-Knowledge Pipeline, acting as a semantic gatekeeper for entity extraction, transformation, and linking steps detailed in Chapter 7. It also provides the semantic context for the Exploration and Application Layer. Because the ontology formally defines that "Parent"is a sub-property of "Relative,"a query for "Relatives"automatically retrieves "Parents"without complex coding. It enables analysts to formulate complex queries and inference models to operate on a logically consistent knowledge base.

However, a conceptual limitation remains. Ontologies excel at explicit reasoning (defined rules) but struggle with implicit reasoning (ambiguity and nuance) without external aid. For example, an ontology can strictly define that "Brother"is a "Relative."Still, it cannot inherently deduce that "two people sharing an address for 10 years likely have a relationship"without probabilistic support. Therefore, the Semantic Foundation is designed to work in tandem with the Composite AI components of the pipeline, described next, to bridge the gap between rigid logical definitions and the fuzzy nature of human behavior.

3.2 THE DATA-TO-KNOWLEDGE PIPELINE

This section describes the process architecture that transforms raw, heterogeneous, and fragmented government data into structured, connected, and actionable knowledge, embodied in the Knowledge Graph. Unlike traditional Extract-Transform-Load models, which are often linear, table-to-table, and content-agnostic, this pipeline is conceived as a cyclical, semantically oriented, and sociotechnical process. Its phases and supporting activities are detailed in Chapter

5.

The design of this pipeline addresses the methodological fragmentation identified in the literature (Chapter 2). Academic literature often treats graph construction as a one-off experiment, whereas government operations require a sustainable, auditable lifecycle. The pipeline addresses the Pragmatic Gap by enforcing a Business Understanding phase at the outset, ensuring alignment with research objectives, and establishing a data governance plan before any technical processing. This approach ensures that graph construction is a strategic tool for answering specific business questions (e.g., "Who is the beneficial owner?"), with full traceability and legal compliance.

The pipeline's operational flow is structured in logical stages. The Knowledge Acquisition phase encompasses the ingestion, preparation, and direct extraction of entities and relationships from structured sources. Next, the Knowledge Improvement stage incorporates an Enrichment and Semantic Inference stage. At this point, the Artificial Intelligence models developed in this thesis, focusing on inferring kinship ties, uncover hidden relationships between environmental offenders, and calculating social proximity (detailed in Chapter 7), are applied to find knowledge that is not explicit in the data but is vital to the investigation. Knowledge Graphs struggle with complex decision-making logic (e.g., "If X and Y, then likely Fraud"). The pipeline integrates the cited Machine Learning models to calculate probabilities, which are then materialized as graph edges or properties (e.g., (:Person)–>(:Company)). This hybrid approach combines the structural clarity of the graph with the predictive power of ML.

It is essential to emphasize that this pipeline does not operate in isolation. The Asset Concealment Domain Ontology serves as the guiding schema across all stages, defining the target structure for entity extraction and ensuring the semantic consistency of the generated knowledge. Furthermore, the Data Governance and continuous knowledge improvement cycles—which are the most important contributions of this process—permeate the entire workflow.

In a legal context, data lineage is paramount. If a graph query suggests an asset seizure, the investigator must be able to trace that conclusion back to the source document. The pipeline is designed to preserve this lineage, creating a provenance graph alongside the knowledge graph and ensuring the auditability of each transformation, from the source to its final representation,

addressing the black box criticism often leveled at AI in government. This lineage tracking is a legal necessity; without it, the intelligence derived from the graph would be inadmissible as evidence, as defense attorneys could challenge the integrity of the data processing chain.

Finally, it is essential to emphasize that the tools and methods outlined in the literature review (Chapter 2) are instrumental at each stage of this pipeline. As presented in Figure 3.1, the pipeline layers abstract the complexity of underlying technologies (NLP, OCR, Graph Databases), allowing the focus to remain on the flow of meaning rather than the flow of bytes.

3.3 THE KNOWLEDGE GRAPH

The Asset Concealment Knowledge Graph (ACKG) embodies the ecosystem's integrated intelligence. This repository represents entities (such as people, companies, real estate, and legal processes) and the complex relationships that unite them explicitly and semantically. It is the end product of the Data-to-Knowledge Pipeline and the concrete instance of the conceptual framework defined by the Domain Ontology.

The choice of Knowledge Graphs as the principal repository of the ecosystem is a deliberate architectural decision to address the architectural fragmentation and functional limitations of Relational Database Management Systems in this specific domain. The decision is grounded in the inherent topology of the asset concealment problem. Financial fraud and asset concealment are topological problems. They involve networks of intermediaries ("straw men"), shell companies, and circular transactions.

A relational database models these as rows in separate tables (Person, Company, Asset). To connect a debtor to a hidden asset through three layers of shell companies requires complex JOIN operations. As the depth of the relationship increases, the performance of RDBMS degrades exponentially. In contrast, a Graph Database stores relationships as first-class citizens (edges). Traversing a relationship is a constant-time operation, allowing investigators to navigate chains of relationships with multiple degrees of separation instantly.

The KG acts as the single source of analytical truth within the ecosystem. It consolidates the data processed and unified by the pipeline, ensuring that each entity and relationship conforms to the ontology's vocabulary and rules. This process transforms siloed data into contextualized

knowledge, where a simple query can reveal, for example, that a debtor (Person A) is a partner in a company (Company B) that owns an asset (Asset C) operated by a close relative (Person D). In a relational system, these facts are disparate records; in the KG, they form a semantic pattern of fraud.

However, a knowledge repository has limited value if it remains inaccessible or unexplored. The KG's true power only manifests itself when analysts and systems effectively consume it to generate actionable intelligence. Furthermore, while the graph excels at topology, it is often less efficient than columnar databases for pure aggregation tasks. Thus, the architecture does not propose replacing all government databases with a Graph, but rather using the Graph as an overlay intelligence layer that references the underlying transactional systems. This hybrid approach ensures scalability while maximizing analytical depth.

3.4 THE ANALYSIS AND VISUALIZATION LAYER

No matter how rich and well-structured, a Knowledge Graph only generates value when its information is consumed to produce actionable intelligence. The analysis and visualization layer serves as the interface between the knowledge repository and end users—human analysts, lawyers, and other computer systems. This layer translates the graph's analytical potential into concrete investigative results. It provides a solution that is operationally useful and integrates with existing workflows, addressing the architectural gap identified in Chapter 2.

A critical conceptual distinction in this ecosystem is the difference between Investigative Intelligence and Legal Evidence. In the domain of asset recovery, these are distinct categories with different standards of rigor. Investigative Intelligence is the road map provided by the graph. It identifies where to look. For example, the graph might show a high probability link between a debtor and a shell company based on shared addresses and phone numbers. This fact is actionable intelligence that guides resource allocation. Legal Evidence is the material admissible in court to prove the link. A screenshot of a graph visualization is rarely sufficient or admissible as evidence due to hearsay rules and the need for original documentation. The court requires the original deed, the bank statement, or the corporate registry filing.

The Analysis Layer is designed to bridge this gap. It allows the analyst to retrieve the source

document that justifies that line and supports the legal strategy by converting the dots and lines of the graph into a substantiated evidentiary package. The system enables the investigator to move from the context of discovery (finding the hidden asset via the graph) to the context of justification (proving ownership via documents).

This layer materializes through three distinct and complementary consumption channels, which ensure the accessibility and usefulness of knowledge for different profiles and needs:

1. Visual Analysis Interfaces (Exploratory): Tools such as Neo4j Browser (NEO4J Inc., 2025) allow investigative analysts to explore the network interactively and intuitively, and leverage human cognition—our ability to spot visual anomalies. It is the primary means of discovering unanticipated patterns, where the graphical representation of relationships enhances human cognition.

2. Declarative Querying (Codified Heuristics): The second channel uses declarative query languages, such as Cypher, which formalize complex investigation heuristics (Chapter 7). Queries in Cypher allow analysts to search for specific, repeatable, and auditable patterns across the entire data set. This channel converts the tacit knowledge of senior investigators into computational logic, ensuring consistent, scalable analysis.

3. Application Programming Interfaces (APIs): Finally, APIs are the third and most strategic channel for corporate integration. APIs expose graph knowledge as a service, allowing other systems to consume it programmatically. In this way, the intelligence about relationship networks can be integrated into process management systems, dashboards, or other analysis tools to enable Decision Intelligence workflows, where the graph runs in the background of standard case management software, flagging risks automatically without requiring the lawyer to be a graph expert.

3.5 THE CYCLE OF GOVERNANCE AND CONTINUOUS IMPROVEMENT

The Governance and Continuous Improvement Cycle ensures the knowledge ecosystem can continually learn, adapt, and evolve. It engages all other components of the framework in an iterative process of refinement and validation. It represents the system's operational in-

telligence, ensuring that the knowledge generated remains accurate, relevant, and compliant with legal and business requirements. This step addresses the methodological fragmentation that often treats the construction of Knowledge Graphs as a linear project (Start → Build → Finish), neglecting the mechanisms required for their sustainability in a dynamic corporate environment.

The cycle operates through two interconnected feedback loops. The first is the Continuous Improvement Loop (Semantic Learning), driven by user interaction with the analysis and visualization layer. When an analyst, while exploring the graph, identifies an information gap or devises a new investigative heuristic, they trigger the Ontology Improvement support activity. This demand leads to refining classes and properties in the ontology or to the inclusion of new data sources to ensure that the ecosystem evolves in direct response to criminals' adversarial adaptation. As the criminals change their tactics, the ontology and graph schema evolve to track them, preventing the system from becoming obsolete.

The second is the Data Governance Mesh, a supporting activity that permeates the entire process. Governance is often viewed as a bureaucratic constraint; in this ecosystem, it is an enabler operationalized by the Knowledge Graph itself. The graph becomes an instrument of modern governance by providing a clear, navigable audit trail. Every edge in the graph has metadata pointing to its origin, satisfying the legal need for provenance. The graph can enforce granular access controls (e.g., masking sensitive personal data for unauthorized users), ensuring compliance with data protection laws such as LGPD/GDPR. This symbiotic relationship—where governance guides the construction of the graph and the graph, in turn, makes governance auditable—fills a critical gap in academic models and is indispensable for application in a public agency. It formalizes the socio-technical collaboration between human actors (analysts, lawyers, knowledge engineers) and technological components, creating a system that learns from experience and improves through use.

3.6 SUMMARY

This chapter presented the reference ecosystem underpinning this thesis's contribution: a Socio-technical Conceptual Framework for discovering hidden assets. The transition from a

linear, layered model to this cyclical, integrated framework is a strategic response to the semantic, methodological, and architectural fragmentation gaps identified in the literature. By positioning the KG as the central, dynamic artifact, fed by a methodological pipeline and governed by a continuous-improvement cycle, this architecture offers an actionable, auditable roadmap for building an investigative intelligence ecosystem.

The interconnection of the components described here—the Semantic Foundation, the Data-to-Knowledge Pipeline, the Knowledge Graph, the analysis and visualization Layer, and the Governance Cycle—is shown in Figure 3.1. The entire process is driven externally by the KG construction process, which details the pipeline steps and implements them through tools belonging to the categories presented (Chapter 2), stacked in the pyramid alongside the process, in the sequence in which they are typically applied in the raw data transformation process.

In the data-to-knowledge pipeline, actors can use different strategies to implement activities based on the tool categories, ranging from the acquisition of private tools to the use of open-source tools, to the development of tools or the creation of AI models that support activities in specific contexts. In the government context, the deliberate choice of explainable, auditable, and legally defensible AI models is based on the requirements imposed on government agencies.

The tools and methods defined and used in each stage are directly proportional to the maturity level of the actor executing the process. In some cases, a lack of maturity may prevent the use of more advanced graph analysis techniques, such as Graph Neural Networks (GNNs). In others, it will prevent the processing of unstructured data. However, this does not mean that integration between actors will be compromised. One actor can consume data from any layer of another, even if the latter has not yet reached the top of the pyramid, as both follow the same process and use the same ontological structure (ACDO), with the same set of entities and relationships.

The actor's maturity also directly influences data selection. Low maturity leads to the use of internal data or open, structured, or semi-structured data. As maturity increases, data obtained through legal sharing and unstructured data, the latter requiring additional processing steps, become part of the available data collection and enrich the results of the investigative process.

It is also important to emphasize that, unlike other methodologies, this thesis advocates separating the Ontology Construction Process from the Knowledge Graph Construction Process.

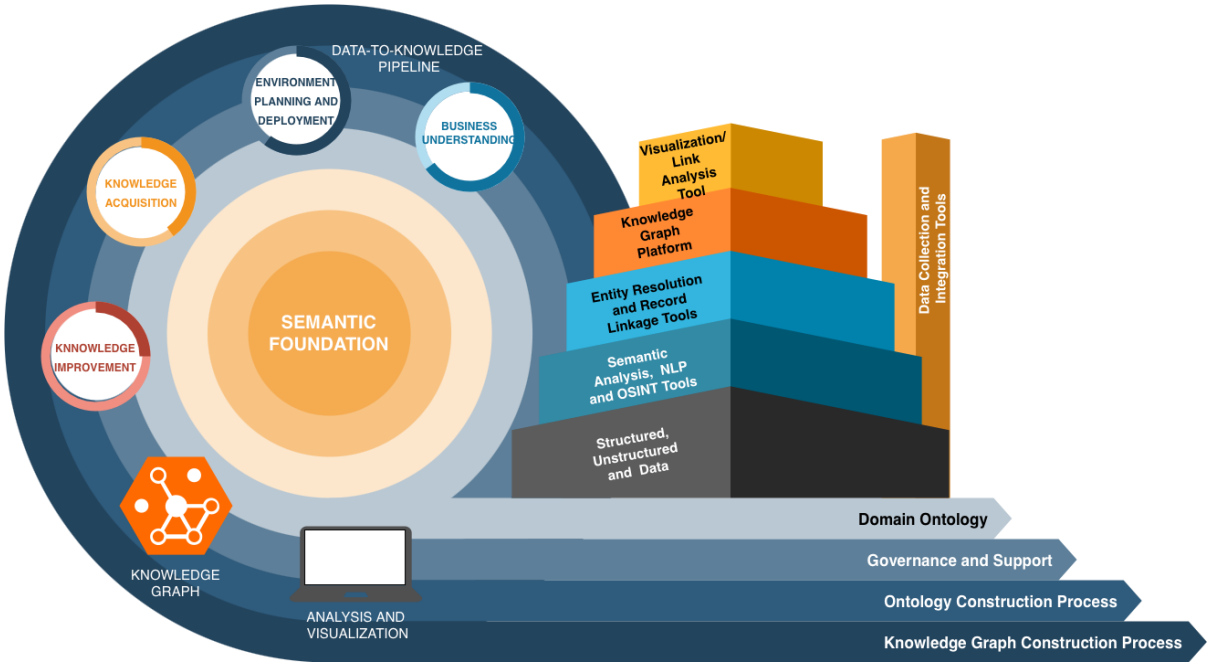

Figure 3.1: Conceptual Framework

For clarity, knowledge modeling (ontology) is a preliminary, conceptual, and iterative process that should guide graph construction, which is primarily a data engineering and instantiation process.

Finally, the conceptual framework detailed in this chapter serves as a map to guide the reader through the thesis's practical and methodological contributions. The subsequent chapters will dissect and implement each component of this architecture. Chapter 4 will detail the construction of the Semantic Foundation, the domain ontology. Chapter 5 will delve deeper into the methodology behind the Data-to-Knowledge Pipeline. Finally, Chapter 7 will present the end-to-end implementation and validation of this ecosystem, demonstrating how this conceptual architecture materializes into a robust and effective technological solution for the Federal Attorney General's Office.

CHAPTER 4

PROPOSAL FOR AN ONTOLOGY BUILDING PROCESS

This chapter presents a detailed proposal for constructing a domain ontology. It is designed to address the specific needs of this research but can be applied to the construction of domain ontologies in general. This proposed process is based on a detailed analysis of the principal methodologies for building ontologies already mentioned in Chapter 2.

The proposal presents an improved version of the seven-step method reviewed in (JIA *et al.*, 2023), incorporating best practices from other methodologies analyzed. Critically, it shifts the paradigm from a purely expert-driven approach to a data-driven strategy. The process introduces new approaches to data integration, real-world data mapping, and ontology alignment. Throughout the following sections, each phase of the process is explained in detail.

4.1 PROPOSED PROCESS

Based on a comprehensive examination of key methodologies for ontology construction discussed in the literature review, a tailored process for building the ontology used in this research was proposed (Figure 4.1). The method proposed by (TORRES *et al.*, 2024) comprises three main phases: specification, implementation, and evaluation. Each of these will be explained in detail in this chapter.

Unlike traditional methodologies, such as those defined in (JIA *et al.*, 2023), which often operate in a waterfall model that relies heavily on extensive initial expert elicitation, this proposal prioritizes the data structure as the primary source of truth. Classic models often suffer from the "HiPPO"(Highest-Paid Person's Opinion) phenomenon, in which subjective expert opinion overrides empirical evidence. By contrast, this process is iterative and anchors conceptual modeling in the actual metadata and data clusters of the target environment, ensuring that the resulting ontology is not an idealized artifact but a pragmatic tool for interoperability.

The proposed process presents a new path for ontology construction, focusing primarily on recognizing relevant external domains in complement to internal data. It can also provide a systematic way of organizing the collected domain data into a preliminary ontology taxonomy. In addition, a strategy is proposed to partially align even existing related ontologies and integrate them with the identified and mapped domain concepts. The following sections give a detailed view of each of these processes. The subsequent chapters will test the creation of an ontology to assert its efficacy and practicality.

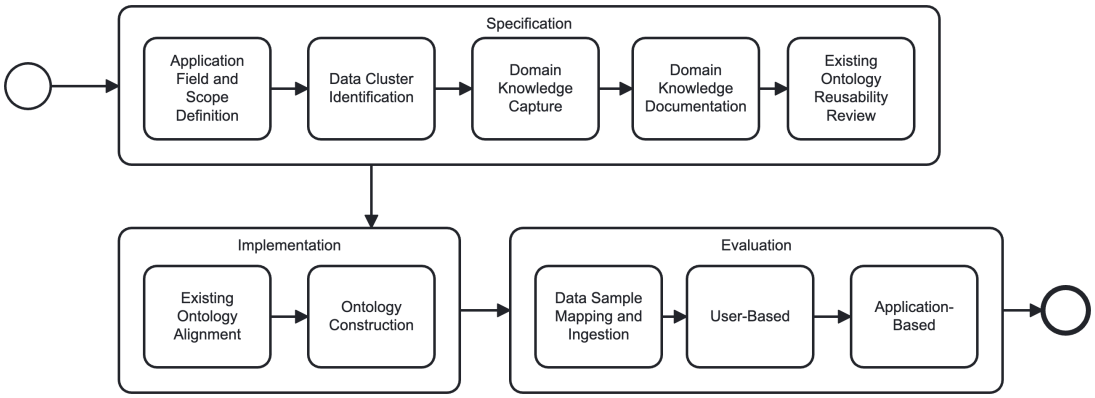

Figure 4.1: Ontology construction process. (TORRES *et al.*, 2024)

4.1.1 Specification Stage

The macro specification process aims to identify the principal elements and requirements needed to build the ontology. It consists of five steps, which will be described in detail below: Application Field and Scope Definition (ASD), Data Cluster Identification (DCI), Domain Knowledge Capture (DKC), Domain Knowledge Documentation (DKD), and Existing Ontology Reusability Review (ORR) (TORRES *et al.*, 2024).

4.1.1.1 Application Field and Scope Definition

The most important part of this activity is identifying the requirements. The initial gathering of requirements would be conducted by contacting stakeholders and experts. The activities involve conducting interviews and surveys to ensure easy understanding of the domain (TORRES *et al.*, 2024).

After that, all these interactions help define the specific needs the ontology should address. In addition, these requirements assess the ontology's effectiveness throughout the process. This phase provides context for ontology development, identifies target users, and builds confidence in the process.

4.1.1.2 Data Cluster Identification

This step's objective is to propose a wider method of domain analysis. In this sense, adding external domain knowledge to the process is one crucial movement. This can be accomplished by including data from outside the organization related to the analyzed domain. The idea behind including external concepts is to incorporate elements that are not part of the internal culture but are essential for understanding the domain. On the one hand, the work becomes more complex, since it is much easier to build an ontology that meets only internal requirements; On the other hand, this would greatly facilitate the integration of data across different companies operating in the same area of activity (TORRES *et al.*, 2024).

This step reinforces the methodology's data-driven character. Instead of relying solely on abstract definitions provided by stakeholders, the engineer must identify "Data Clusters"—groupings of actual datasets, schemas, and legacy databases. This empirical approach mitigates the risk of modeling non-existent concepts and ensures the ontology is grounded in the reality of the available information systems.

Rather than delving into the specific functionalities of each internal system, these systems should be categorized by general areas of interest or themes. The establishment of categories enables a broader search framework that can connect to one or more topics, rather than being restricted to the context of a single organization.

4.1.1.3 Domain Knowledge Capture

This step focuses on identifying data and metadata related to the domain. It aims to expand the search scope beyond the organization developing the ontology. Therefore, internal and external databases relevant to the problem being analyzed should be mapped. At this stage, accessing the actual data within the databases is unnecessary. Instead, the metadata

includes information about the stored attributes, even if only partially (TORRES *et al.*, 2024).

Identifying the actors involved in the process is a requirement that needs to be addressed before the search begins. For example, if the domain under analysis involves data managed by governments, relevant countries should be identified as potential actors. It ensures that the search process can be organized and guided for each defined category, with a focus on these participating actors.

With the categories established in the previous step and the actors and databases/metadata related to the domain identified, the final activity is to map the possible classes, attributes, and relationships associated with the domain for each data category. Table 4.1 provides an example of mapping and documenting possible attributes, entities, and relationships (TORRES *et al.*, 2024).

4.1.1.4 Domain Knowledge Documentation

During the Domain Knowledge Capture phase, concepts, attributes, and domain relationships should be identified and recorded. These results, which are initially organized by category in DKC, must later be combined into a single, cohesive structure during the Domain Knowledge Documentation step.

This approach ensures that similar concepts across different categories, along with their specific properties and relationships, are consolidated into a unified representation. The primary objective is to develop a preliminary ontology based solely on the knowledge mapped from internal and external data sources (TORRES *et al.*, 2024). In many of the methodologies analyzed, this deliverable often represents the final taxonomy structure or the relationships among concepts for the ontology being developed.

Table 4.1: Proposed Table of Entities, Attributes, and Relationships identified (TORRES *et al.*, 2024).

Attributes	**Entities and Relationships**
• Class_1: attribute_1, attribute_2 • Class_2: attribute_3, attribute_4	$Class\_1 \xrightarrow{RELATIONSHIP} Class\_2$

4.1.1.5 Existing Ontology Reusability Review

Evaluating the reusability of existing ontologies involves identifying domain-related ontologies and determining whether they can be reused (TORRES *et al.*, 2024). In this methodology, reusability is treated as a fundamental architectural decision. The reinvention of the wheel is explicitly discouraged. This step mandates searching for Ontology Design Patterns (ODPs) and foundational ontologies to provide a robust backbone for the new model.

A key aspect of this methodology is leveraging ontologies that partially address the domain in question. If no existing ontology fully captures the knowledge needed for a specific domain, such as asset concealment, high-level ontologies or ontologies from related fields can be used to complement the development of a vocabulary of classes, attributes, and relationships for the new ontology. As a result of this step, a collection of relevant existing ontologies, including high-level ones, is expected.

For example, asset concealment is a specialized area that focuses on mapping individuals, their assets, and the entities and relationships involved in concealing those assets (TORRES *et al.*, 2024). Although there isn't a dedicated ontology for this domain, many overlapping concepts and relations related to finance, social relationships, and so on, such as those for family and friends, do exist. Therefore, surveying existing ontologies in these domains is warranted to allow the reuse of relevant concepts. Furthermore, for each ontology analyzed, concepts and attributes that align with the domain under study should be selected, as these ontologies often contain thousands of mapped concepts.

4.1.2 Implementation Stage

The implementation stage builds upon the specifications outlined in the previous phase. It begins with the Existing Ontology Alignment (EOA) activity, which focuses on unifying relevant concepts or classes related to the domain, as identified in the previous ORR step (TORRES *et al.*, 2024). It is essential because multiple ontologies developed for similar domains often contain overlapping or similar concepts, but may use different terminologies or structures.

The activity of aligning these concepts involves identifying correspondences between entities (such as classes or properties) across different ontologies to ensure interoperability and consis-

tency among systems. These definitions must be consolidated into a unified representation of each concept. Additionally, hierarchical relationships are established during this step.

The outcome of the EOA activity is a domain ontology that defines classes and relationships based on existing ontologies. Following this, the ontology construction phase (OCO) focuses on creating the final ontology by merging the aligned ontologies with the concepts and relationships identified during the Domain Knowledge Capture (DKC) and Domain Knowledge Documentation steps.

4.1.2.1 Existing Ontology Alignment

The existing domain ontologies contain concepts that ought to be identified and aligned. These are aligned into relationships for ideas that may be similar but carry different identifiers across various ontologies (TORRES *et al.*, 2024). This process would help to enable better unification in the application of terminologies within the domain under consideration and would lessen divergent meanings attributed to the concepts. Correct identification and correlation among these concepts are necessary to ensure consistent data representation and interpretability within the domain being studied.

As illustrated in 4.2, the alignment process starts with the formal definition of each concept, where each ontology maps classes associated with the defined concept. It is important to note that only concepts directly related to the domain should be included in this process.

Table 4.2: Proposed Table for alignment of ontology classes (TORRES *et al.*, 2024)

Concept Description	**ONTO-1**	**ONTO-2**	**ONTO-3**
A corporate or similar institution.	Organization	Organization	Company
Modern man, the only remaining species of the Homo genus.	Person	Human	Person

In the sequence, a careful analysis of their related properties is needed to map the relations among them. This analysis will help to illustrate the complex interrelation among the ideas. It will ensure that the ontology accurately represents the interactions within the domain (TORRES *et al.*, 2024). Mapping out these relations generates a more comprehensive framework that

identifies the ideas and their interdependencies, thereby improving their utility. An example of relationship alignment has been provided in 4.3.

Table 4.3: Proposed Table to Alignment of ontology relationships (TORRES *et al.*, 2024)

Feature	**ONTO-1**	**ONTO-2**	**ONTO-3**
Parentage	isParentOf isFatherOf	- MFatherOf	hasChild / hasParent -

4.1.2.2 Ontology Construction

This phase focuses on integrating the aligned classes, attributes, and relationships from existing ontologies with those identified during the Domain Knowledge Documentation stage. The knowledge engineering team and domain analysts must carefully select the most relevant concepts and relationships from the mapped data to build the final ontology (TORRES *et al.*, 2024). The completed ontology should be structured using a widely accepted format. In this study, we recommend using the OWL or RDF formats for representation.

4.1.3 Evaluation Stage

Ontology evaluation assures that the ontology can accurately serve its intended purposes and represent its domain knowledge. To reach this goal, many elements are evaluated, including the correctness, completeness, consistency, and usability of ontologies (TORRES *et al.*, 2024).

This phase identifies potential issues that might arise, thus improving the quality. It also ensures that the ontology assists adequately with reasoning, querying, and interoperability. Based on these evaluations, the ontological structure is fine-tuned where required.

Often, different algorithms and evaluation methods are utilized for ontology quality evaluation. The evaluation begins by loading a small amount of reference data into the ontology to verify that all relevant concepts and associations are captured. In a sequence, a user-based and an application-based evaluation are employed. Through this evaluation process, such errors, gaps, and the most pressing issues will be identified and given importance. These results thoroughly analyze the ontology's strengths and weaknesses.

4.1.3.1 Data Sample Mapping and Ingestion

This phase's main idea is to confirm that the ontology correctly maps concepts and that all relevant attributes are appropriately assigned to their respective classes and corresponding relationships (TORRES *et al.*, 2024).

At this stage, a subset of the data is incorporated into the implemented ontology as a preliminary test to determine whether it can accommodate the full range of attributes intended for future use. If the ontology is well-structured, all key attributes and relationships within the sample should be appropriately represented within the defined classes and relationships.

After this phase's execution, the ontology must be sufficiently adaptable to support real data integration. In this way, inconsistencies and missing elements must be corrected before large-scale data incorporation.

4.1.3.2 User-based Evaluation

Correctness is a critical aspect of ontology evaluation, focusing on the ontology's likeness to the domain it is intended to model. The task of correctness also requires checking the concepts, relationships, and attributes stated in the ontology against real-world entities and their interactions. Completeness is another crucial aspect that guarantees the ontology comprises all the most relevant concepts and relationships from a domain. The ontology is expected not to omit any essential entities or interactions that could be fundamental for reasoning and querying. Completeness can be tested by comparing the ontology against the domain requirements to see whether anything has already been ignored.

User-based evaluation allows domain experts and stakeholders identified during the specification phase to review and instantiate the ontology, giving feedback on correctness, completeness, and usability (TORRES *et al.*, 2024). The foremost aim is to determine whether the ontology covers the entire domain of interest, whether all concepts, relations, and attributes necessary to the relevant application must be checked, and whether the detailed ontology levels fit. This element should also be examined against each high-level domain requirement.

Subsequent revisions are necessary to address the identified problems. An analysis is carried out, which will include all amendments that could be made to address inconsistencies and

errors, improve coverage-skimming by identifying concepts and relationships that seem missing, improve definitions, and clarify words that could be called into question for soundness and general usability of the ontology at hand (TORRES *et al.*, 2024).

4.1.3.3 Application-based Evaluation

In a practical, application-based assessment, the ontology is implemented in a concrete application—e.g., a knowledge graph—and its performance is evaluated based on its ability to facilitate the application's operations. This type of assessment aims to evaluate the ontology's fitness, performance, and general effectiveness in the real world. In contrast to theoretical evaluation, which focuses mainly on the internal organization and logical coherence of the ontology, application-based evaluation emphasizes practical usability and applicability.

A significant part of this analysis is the principle of consistency, which means the ontology should not contain contradictory definitions or rules. To determine this, reasoning engines are often used to detect and resolve logical inconsistencies, ensuring that automated reasoning can proceed without generating contradictions or errors.

This assessment checks how the ontology harmoniously integrates with other systems and data repositories. Domain experts must be included in the validation process to provide seamless interoperability. The outcome should demonstrate that the ontology can communicate effectively with other systems.

4.1.4 Methodological Limitations and Critical Analysis

While the data-driven approach proposed in this chapter offers significant advantages in scalability and empirical grounding, it is necessary to acknowledge its methodological limitations. The reliance on existing data clusters introduces the risk of the "Garbage In, Garbage Out"phenomenon, in which poor data quality, inconsistencies, or historical biases in the source databases are inadvertently encoded into the ontology structure.

Furthermore, automatically or semi-automatically induced ontologies may lack the semantic richness of those handcrafted by philosophers or senior experts. This methodology mitigates

these risks by maintaining a "Human-in-the-loop"approach during the Evaluation Stage, but the potential for algorithmic bias remains a critical concern. The process assumes that the available data is a representative proxy for the domain reality, which may not always be true in cases of incomplete digitalization. Therefore, this process should be viewed as a complement to, rather than a complete replacement for, expert judgment.

4.2 SUMMARY

This chapter presented the first methodological contribution of the thesis: a generic process for constructing domain ontologies. The proposed methodology is structured into three stages—Specification, Implementation, and Evaluation—and differs from traditional approaches by its data-driven emphasis, which anchors the modeling in the analysis of real data clusters. This innovative approach ensures that the resulting ontology is pragmatic, directly applicable, and less susceptible to the abstraction gaps common in purely top-down methodologies.

This section details the process's validation, focusing on its ability to support the alignment of existing ontologies and its robustness in real-world applications. With this methodology for establishing the Semantic Foundation, the next chapter will present the second methodological contribution: a structured process for constructing the Knowledge Graph, which will consume the ontology produced by this method.

CHAPTER 5

PROPOSAL FOR A KNOWLEDGE GRAPH BUILDING PROCESS

This chapter proposes a methodological approach to constructing a Knowledge Graph that addresses limitations identified in the processes analyzed in the literature. Initially, the chapter reviews the limitations of the knowledge graph development process proposed by (TAMAŠAUSKAITĖ; GROTH, 2023), highlighting issues such as the lack of structured participation from business stakeholders, insufficient definition of roles and responsibilities among technical profiles, and inadequate treatment of ontology development as a distinct, iterative process.

Beyond operational steps, this proposal aims to fill a significant gap in the literature: the lack of methodologies that reconcile the flexibility of semantic technologies with the rigid requirements of corporate environments. Most existing models focus on the technical extraction of triples but fail to provide a framework that ensures data durability, legal compliance, and business alignment. Therefore, the scientific motivation for this new process is to deliver a corporate-ready framework that treats the KG as a dynamic knowledge ecosystem governed by explicit semantic rules.

The proposed process requires continuous stakeholder involvement and a structured approach to ontology modeling. It also includes data governance practices and proposes a flexible execution model for iterative refinements. The idea is to redefine the execution cycle by structuring key phases: business understanding, technological environment planning, knowledge acquisition, knowledge improvement, and deployment. Furthermore, it introduces governance and ontological refinement as continuous processes throughout the KG lifecycle.

5.1 CONTEXTUALIZATION

The methodology initially proposed for constructing the knowledge graph within the scope of this work was based on the research presented in the literature review section cite tamav-

sauskaite2023defining. The premise was that the authors had successfully captured the main stages of the development process by analyzing 57 relevant articles.

However, during the practical execution of the activities, a series of issues that were not present in the process in question and needed to be addressed, as well as activities present in the original process that required adjustment, were noticed. The authors warned of the need to test the process in a real environment. They pointed out gaps in some aspects, such as the need to have a moment to define the technological architecture, although this is not explicitly stated in the process. As the applied changes significantly modified the original process, a new process for Knowledge Graph Building was proposed. This new approach shifts the focus from a purely algorithmic task to a socio-technical one, where the alignment between the organizational "Business Understanding"and the technical "Knowledge Acquisition"is the primary driver of success.

5.2 ISSUES IDENTIFIED

One of the main problems observed in the analyzed process is the predominance of an essentially technical perspective. The steps are presented almost exclusively in terms of methods and tools for extracting entities, relationships, and attributes, as well as graph integration and maintenance. Although such activities are fundamental, it is unclear when and how business stakeholders — experts with detailed knowledge of the domain in which the graph will be applied — should be involved.

In practice, the lack of formalized, structured participation by these stakeholders led to several challenges. Since they were not allocated full-time to the project and had reduced schedules, the need to consult them at various stages of the development cycle created scheduling difficulties. This absence especially impacted the requirements-gathering activity, the definition of business objectives that justify the construction of the graph, and the entity and relationship validation process, creating gaps that could have been avoided with a more consistent and organized presence of the business team. From an analytical standpoint, the absence of domain experts creates a semantic gap where the extracted graph may be technically correct but business-irrelevant.

In line with the need to clearly define the business team's participation throughout the process, a similar situation was observed for the other profiles involved. In general, knowledge graph projects rely on several technical profiles, but the process analyzed does not clearly separate at which stage each profile should act as a protagonist or collaborator. In the Extract Knowledge, Process Knowledge, and Construct the Knowledge Graph phases, practically the entire team was mobilized, generating inefficiencies. To improve execution, it would be essential to structure the process so that each step highlights the profile that plays a key role, clearly defining when domain experts are most needed, when data engineering activities occur, and when ontological validation takes place.

Another critical point concerns how the current process treats the creation of ontologies as just another step in the construction of the knowledge graph. Analytically, it is vital to distinguish the ontology (the T-Box or schema level) from the Knowledge Graph (the A-Box or instance level). While the ontology defines the logic and constraints, the KG represents the grounded data by treating ontological development as a direct substep of the main flow, undersized resources (time, personnel, and tools), and confusing teams, who are forced to conduct conceptual modeling while simultaneously performing subsequent activities, such as knowledge extraction and data integration. In complex domains, such as asset concealment, the ontology requires careful iterations to ensure conceptual coherence. For this reason, it is recommended to treat ontological development as an autonomous process, even if connected to the knowledge graph project.

Data Governance and Quality is another point to be addressed. The process does not explicitly address this topic at any stage, and this was one of the most critical points encountered in practical execution. In practice, it was also observed that, although there is talk of "Processing" and "Integrating" knowledge, there is a lack of precise data governance mechanisms, such as versioning, traceability (lineage/provenance), quality control, and access policies. Such elements become crucial in corporate scenarios, where it is vital to ensure compliance with legal standards (General Data Protection Regulation - GDPR) and to maintain data integrity and security. The corporate-ready aspect of a KG depends entirely on these non-functional requirements, which are often overlooked in academic prototypes but are essential for scientific contributions in applied computer science.

Finally, although storage alternatives are mentioned (RDF triple stores, graph databases, key-value stores), there is no formal moment to decide on the technological architecture most appropriate to the use case. This decision is especially relevant when the organization is building its first knowledge graph, as was the case in the experiment discussed here.

5.3 PROPOSED PROCESS

Based on the previous critical analysis, we propose a reformulation of the knowledge graph construction process to reinforce the participation of business experts; treat the creation of ontologies as an independent but integrated process with a specific stage of continuous improvement; make the division of professional profiles in each stage more clearly explicit, address the necessity of data governance and compliance with legislation, and create a process for discussing technological infrastructure, among others minor issues.

A structural change was also proposed in the process execution cycle. While the original process is typically presented linearly, with sequential steps and limited return links, the proposed model supports large groups of activities and allows free flow between these subgroups during execution. Although it is expected that the process starts with the Business Understanding stage and ends with the Knowledge Graph Use and Evaluation stage, transitions between states are allowed so that return flows can occur whenever necessary. In addition, two support activities - Ontology Improvement and Data Governance- remain in execution throughout the process cycle. The proposed process is presented in Figure 5.1.

5.3.1 Business Understanding

It involves business experts who know the application domain or sector to validate if the initiative aligns with the organization's strategic needs and can generate tangible value for end users. It is also a time to align expectations between the technical team and the team of business experts to ensure a common understanding of the capabilities and limitations of using the knowledge graph in the organizational context. These elements will reduce the risk of dissatisfaction or subsequent rework.

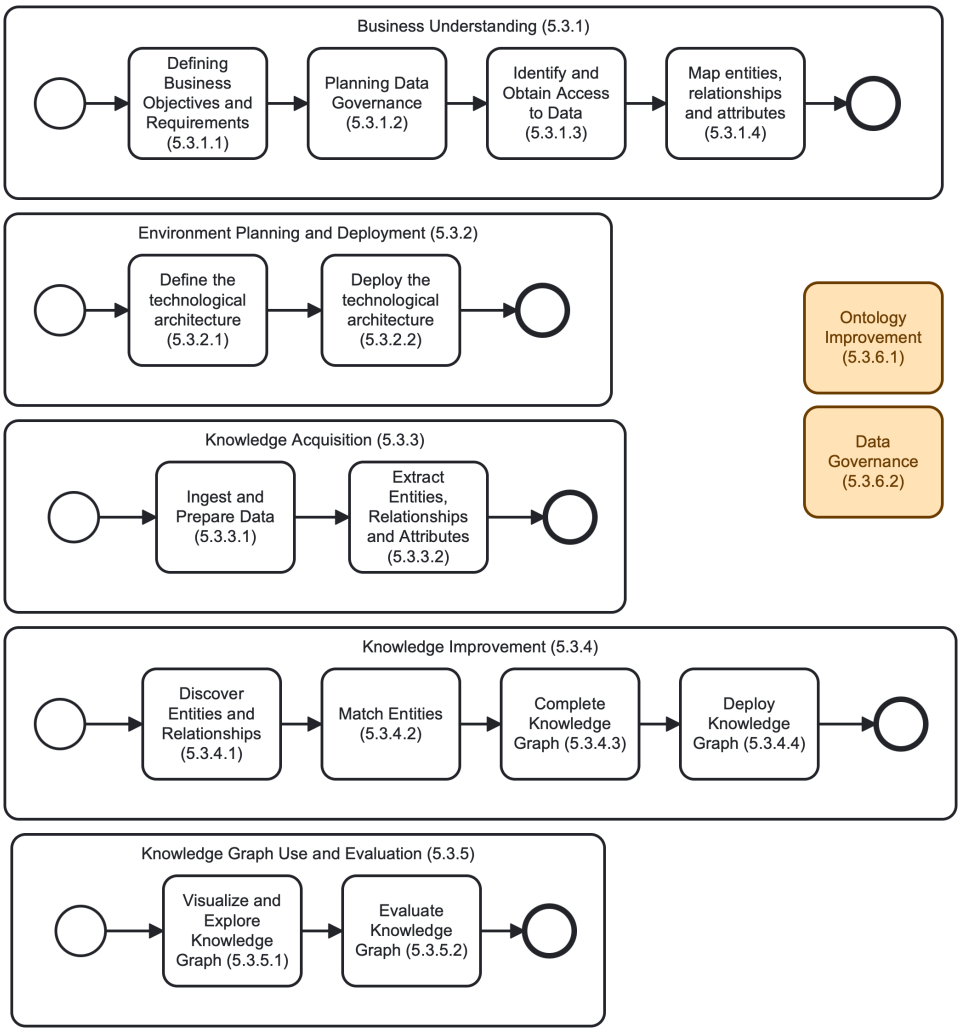

Figure 5.1: Knowledge Graph Build Process

5.3.1.1 Defining Business Objectives and Requirements

In this first phase, the purpose of the knowledge graph is established at a strategic level, determining its value to the organization and end users. The idea is to identify the main problems or opportunities to be addressed, for example, improving decision-making processes, discovering insights in complex data, or supporting personalized services. To this end, senior management, project managers, and domain experts are involved, ensuring that the graph's goals are aligned with corporate priorities.

This stage results in a Requirements Document that describes the business objectives, success metrics (performance indicators), and how the graph will support internal activities and processes. It is important to note that, in this context, the ontology—even if it already exists—is not yet thoroughly analyzed. The emphasis is on understanding which organizational questions

the graph should answer or which improvements it should enable.

5.3.1.2 Planning Data Governance

Once the objectives have been clarified, a Data and Knowledge Governance Plan is drawn up. Governance is a methodological pillar that ensures the KG's reliability. It includes the definition of roles and responsibilities, identifying who will be responsible for curating, versioning, and validating the graph content. The team should also define quality and versioning policies through the specification of documentation standards (metadata), version control (of both data and ontology), and rules for dealing with discrepancies or duplicate information. In a scientific context, governance provides the provenance necessary for reproducibility and auditability.

In this phase, mechanisms to comply with legislation such as GDPR, involving anonymization, consent records, and restrictions on access to sensitive data, should be adopted. Audits and monitoring are also planned, defining how the organization will track changes, who will have read and write permissions, and how any security breaches will be handled.

The primary outcome of this step is a plan that standardizes data management and the ontology. This document will guide future decisions regarding data ingestion, validation, and updates.

5.3.1.3 Identify and Obtain Access to Data

Once the governance bases have been defined, the next step is to identify and obtain the administrative permissions required to access the data sources that will feed the graph. This step maps internal repositories (relational databases, files in various formats, corporate APIs) and external data (public-domain data, specialized sources, or partnerships) that may enrich knowledge. The team must also assess access restrictions imposed by issues involving confidentiality, privacy, or license of use to ensure that all legal and governance requirements are being met. At this time, the physical acquisition of data sets via ETL pipelines is not yet being addressed; it is only about obtaining access rights.

Therefore, although this phase focuses on collecting and mapping sources, there is potential

for feedback on governance planning if additional security requirements or compliance issues arise. As a final product, an inventory of data sources with clear access routes and established reliability criteria are obtained, allowing the conceptual mapping stage, or adjustments to the ontology, to occur consistently.

5.3.1.4 Map Entities, Relationships, and Attributes

After understanding the available data and establishing governance, the domain entities, relationships, and attributes are mapped or aligned with the ontology that the project already has as a starting point. It is the moment where the abstract concepts of the ontology are grounded in the concrete reality of available databases. Knowledge engineers and business experts must perform a comparative analysis of the current ontology and the data that will be incorporated. If new entities, relationships, or relevant properties are identified, the ontology can be reviewed and updated in accordance with the versioning process defined in governance. Domain experts are called upon not just to validate but to co-design the mapping, ensuring that the semantic translation of a database column to an ontological class maintains the original business intent.

The team must duly record every change or expansion of the ontological model, associating the change history with the data governance rules, for example, who approved the creation of a particular property or class. The expected result is a revised ontology that accurately reflects the domain and available data sources. This model will guide the ingestion, extraction, reconciliation, and graph use stages, ensuring semantic consistency and alignment with governance policies and business objectives.

5.3.2 Environment Planning and Deployment

After the Business Understanding phase — in which business objectives, requirements, and scope are defined — and before Knowledge Acquisition itself, it becomes necessary to plan and make available the technological infrastructure that will serve as the basis for the knowledge graph. This step aims to ensure that the architecture, tools, and computing resources selected fully align with the mapped needs and can support the following process activities. Additionally,

the chosen tools must be deployed in the several environments defined in the data governance plan, considering scalability, performance, and operational cost requirements.

5.3.2.1 Define the technological architecture

In the first activity, the team must evaluate various tools for creating and deploying the knowledge graph. A non-extensive list consists of options for storing raw data, intermediate formats, and the final knowledge graph. ETL tools and tools for creating and evaluating ML models should also be assessed.

The evaluation process must consider not only technical criteria but also the organization's reality. While the chosen tools must be appropriate to the expected data volume, query complexity, and scalability requirements, they cannot ignore the organization's existing technological infrastructure and possible budget constraints, among other factors.

5.3.2.2 Deploy the Technological Environment

Once the architecture definition is complete, it is time to provision servers, configure databases, and establish the routines required for the systems to function correctly. The deployment can be executed using different models, such as cloud infrastructure, on-premises, or a hybrid model, always observing the requirements identified in the previous step and the legal requirements defined in the Governance Plan.

At this point, performance and scalability tests are performed, adjusting network, storage, and processing parameters as needed. The environment must also be integrated with governance tools to enable audits, data lineage records, and the application of access policies. The result is a stable technological ecosystem ready to ingest data, robustly supporting the subsequent phases of Knowledge Acquisition and other stages of the knowledge graph construction process.

5.3.3 Knowledge Acquisition

The knowledge acquisition phase aims to create structured and unstructured data ingestion pipelines for the ecosystem's internal environment. Each data type is given appropriate tre-

atment before ingestion. Specifically, regarding unstructured data, it is essential to highlight the need to automate the identification of concepts and relationships hidden in free text in the next step.

After the data is incorporated, only structured or semi-structured datasets are eligible for direct extraction of entities, relationships, and attributes. It includes extraction pipelines built without ML techniques.

5.3.3.1 Ingest and Prepare Data

The main objective is to locate, collect, and prepare the various data sources that will feed the knowledge graph. Based on the mapping performed in the previous phases, the team identifies where the relevant data is located, including structured repositories, semi-structured sources, and unstructured sources. In some scenarios, the process includes web crawling techniques to collect publicly available information on the Internet.

Data pre-processing involves tasks such as cleaning, normalization, and transformation in which duplicate records, inconsistent values, or incomplete entries are removed. Normalization involves unifying formats and nomenclatures, ensuring the team can link the data to the defined evolving ontology. Data Engineers can also adapt structures and formats during transformation to suit ingestion pipelines or specific tools. These procedures are guided by the previously defined governance policies, which ensure compliance with quality, versioning, and security indicators.

This stage is completed by verifying whether the sources meet minimum requirements for completeness, consistency, and compliance. Data engineers and ETL analysts supervise the execution of this verification and are also responsible for documenting what was collected and processed. In addition, any significant anomalies identified can lead to adjustments to the ontology or governance planning, including the need to review access definitions or metadata standards.

Finally, pre-processed and documented data is obtained and ready for use in the subsequent phase. It will facilitate the detection of entities and relationships with greater precision and reduce the incidence of future errors.

5.3.3.2 Extract Entities, Relationships, and Attributes

The primary objective of this step is to directly extract entities, relationships, and attributes that were previously mapped between the data and the ontology. The premise is that the data is organized in a sufficiently transparent and standardized way so that a data engineer can directly associate it with the concepts and properties of the semantic model without the use of machine learning techniques or other advanced Natural Language Processing methods (such as NER, Entity Matching, or disambiguation algorithms).

This step only deals with structured or semi-structured data, for example, in which each table or set of columns corresponds to a specific entity in the ontology. At the same time, foreign keys or equivalent identifiers represent potential relationships. Similarly, explicitly defined text or numeric fields can be mapped directly as attributes without additional inference. The team must formalize and document the conversion rules used to transform each field's values into instances or edges consistent with the ontology, ensuring traceability and facilitating future revisions.

Concept and attribute mapping in structured databases is a relatively simple task. Access to the data dictionary of tables and attributes, or to people with knowledge of the domain, is necessary to identify which table characteristics contain the concepts to be mapped. The difficulty with the concepts comes precisely from the need to integrate the same entity, already mapped and stored in a dispersed manner in different tables in different databases. As a result, the elements are not yet ready to feed the KG and must be stored in a temporary storage medium until the matching process is carried out in the following steps.

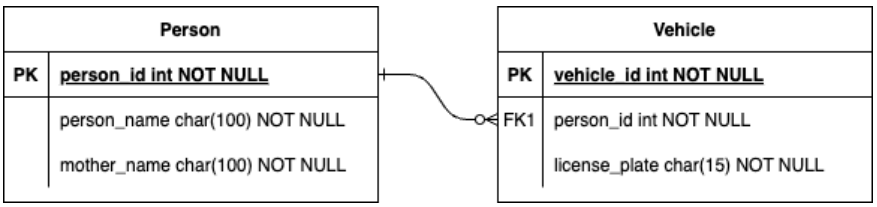

Figure 5.2: Mapping relationship in Structured data.

Regarding relationships, the complexity is greater, except when relationships are created from attributes with unique identifiers. Take, as an example, the creation of a relationship between a specific person, registered in the database with a unique identifier, and a vehicle-

type asset, whose record is stored in a different database, which stores the record but which contains owner information defined through a unique identifier, is almost automatic (Figure 5.2)

In the end, this step generates a set of entities and relationships that are directly mapped to the ontology and stored in an intermediate storage, ready for processing in an entity-matching or record-linkage process.

5.3.4 Knowledge Improvement

This stage comprises four fundamental subprocesses that, in an integrated manner, handle the discovery of entities in unstructured data, the reconciliation of duplicate or conflicting entities, the enrichment of the knowledge graph, and the final deployment in a production environment. Initially, techniques such as Natural Language Processing and Machine Learning algorithms are applied to extract names of people, organizations, products, and other relevant concepts from unstructured data, ensuring that this information scattered across text reports, PDF documents, and social media content is identified and organized.

Next, records describing the same object are compared and unified, avoiding duplication and ensuring graph consistency, since the same entity should not appear under multiple identifiers. Subsequent consolidation aims to expand and complete the represented knowledge through semantic inferences based on ontology rules or by increasing previously unexplored relationships and attributes.

Finally, the graph is deployed in a test or production environment, with attention to scalability, availability, and security, as well as the configuration of the necessary query or visualization tools, so users can effectively use the knowledge and integrate it across different corporate systems.

5.3.4.1 Discover Entities and Relationships

The entity and relationship discovery stage focuses on exploring the available data — structured or unstructured — to automatically identify possible instances, classes, properties, and

links relevant to the domain. In this phase, Named Entity Recognition techniques, relationship recognition algorithms, and, when necessary, machine learning and Natural Language Processing methods are applied. The goal is to extract information not explicitly or directly mapped in the dataset.

Extracting classes from unstructured text is a complex task. It involves identifying and categorizing entities in unstructured text into predefined classes. One challenge is ambiguity, as words can have multiple meanings depending on context. Another issue arises from the variability in language use when synonyms or different words express the same concept.

Technical and computational challenges also play a significant role since developing algorithms capable of handling the intricacies of natural language is inherently complex. Balancing accuracy with computational efficiency is a constant challenge in NLP. Additionally, training models for class extraction often requires large annotated datasets, which may be sparse or unavailable in specific domains or languages. Overcoming data sparsity requires techniques such as transfer learning, data augmentation, and the use of unannotated data.

This process obtains an additional set of entities, attributes, and connections that must be associated with the classes and properties previously defined in the ontology that were not mapped in the previous step. The team members responsible for this activity include data scientists, NLP specialists, and knowledge engineers, who are responsible for both the design and execution of the extraction models and the semantic verification of what has been discovered. Domain experts must be involved to validate the results found to ensure that the findings make sense in the organizational context.

5.3.4.2 Match Entities

Once entities and relationships have been identified through automated analysis or direct discovery, the need arises to reconcile redundant or overlapping records. The primary objective of this stage is to analyze entities to avoid duplications, such as two instances that refer to the same customer or product, and ensure consistency in the knowledge repository. Entity resolution (record linkage) techniques are applied to achieve this goal.

In this process, data engineers and knowledge engineers work together to define the matching

rules or algorithms, considering criteria such as phonetic similarity, string approximation, provenance metadata, and semantic context provided by the ontology. The governance team can also verify that merges or deduplications comply with institutional policies, especially when there are legal or audit requirements.

A typical data-matching process, as described in (CHRISTEN, 2012) and (KEJRIWAL, 2019), involves two main high-level steps. The first is blocking, which clusters approximately similar entities into overlapping blocks, generating candidate record pairs likely to correspond to the same entity, to reduce the quadratic complexity of the matching process. In the second step, these candidate pairs are compared using various functions and sorted into matches, non-matches, and potential matches, with the latter requiring manual review.

Various metrics can be used to quantify the similarity between records, such as Hamming distance (HAMMING, 1950), Levenshtein (LEVENSHTEIN, 1966), Damerau-Levenshtein (DAMERAU, 1964), Jaro similarity (JARO, 1989), Optimal String Alignment (HERRANZ *et al.*, 2011), among others. The result is a set of entities, where each instance corresponds to a unique object in the domain, avoiding duplication and inconsistency.

5.3.4.3 Complete Knowledge Graph

In the third stage, the focus is on completing and consolidating the knowledge graph to integrate all extracted and reconciled data. Here, it is common to apply semantic reasoning and inferential rules to discover new relationships or attributes that were not made explicit and to check for remaining gaps or inconsistencies. When relevant, automatic validation mechanisms, e.g., Shapes Constraint Language (SHACL) or integrity rules, help detect violations of the ontology or defined governance policies.

As the information is enriched and aligned with the ontology model, the graph reaches a coherent state in which each element (entity, relationship, or attribute) is correctly connected to the others, according to the established semantics. In this phase, the participation of domain experts and knowledge engineers is crucial to ensure that the inferences make practical sense and that any revisions to the ontology are controlled.

Graph-based algorithms also contribute to knowledge graph completion by leveraging the

graph's structural properties. Algorithms such as Graph Neural Networks have gained importance due to their ability to capture complex relationships and dependencies within graphs (WU *et al.*, 2020). Rule-based methods are also helpful in KG completion (GALÁRRAGA *et al.*, 2013). These methods involve deriving logical rules from the existing graph to infer new relationships.

At the end of this step, a complete knowledge graph is obtained, consistent with the ontology, and ready for deployment, visualization, or integration with other applications. This final product is the foundation for knowledge discovery applications, advanced data analysis, and support for the business activities that motivated the project.

5.3.4.4 Deploy Knowledge Graph

The first step is to insert the information (e.g., triples or node-edges) into a storage system that supports the chosen model, whether a triplestore (for RDF data), a graph database (such as Neo4j), or another repository optimized for the relationships defined in the previous step. Next, indexes and optimizations are configured, adjusting parameters to enable fast, efficient searches, especially when complex queries or a high volume of simultaneous accesses are expected.

Software architects, developers, and Database Administrators specialized in graph databases are essential to ensuring that the structure, design, and parameterization meet performance and reliability requirements. In addition, query tools or APIs are established, such as SPARQL (for RDF), Cypher (for graph databases), or GraphQL, so that users and other systems can interact with the graph consistently.

In the end, the KG is deployed in a production or test environment and becomes accessible to internal or external applications according to the defined governance policies. This milestone enables both exploratory analysis and the creation of solutions that rely on complex reasoning, personalized recommendations, or integrations with business processes. In this way, the organization begins to reap the strategic benefits of an integrated knowledge repository ready for queries, analysis, and continuous evolution.

5.3.5 Knowledge Graph Use and Evaluation

After the construction and deployment phases, the knowledge graph transitions from a purely technical artifact to a strategic resource supporting various organizational needs. At this stage, the graph is no longer confined to development or testing environments; it becomes accessible to diverse user profiles and integrated into day-to-day workflows or analytical processes. It allows the organization to obtain real value from the consolidated data.

From a business perspective, the knowledge graph must meet the defined needs. As teams explore entities, relationships, and attributes in real-world scenarios, potential refinements to both data and ontology often come to light. These tweaks involve redesigning the coverage for some classes, adding new properties based on requests from domain experts, or even expanding the dataset to include external data sources that were out of scope at the start.

Monitoring response times and load behavior under different usage patterns also helps verify whether the chosen architecture meets the organization's standards and can handle future growth. In parallel, measuring data quality—completeness, consistency, and correct alignment with the ontology—ensures that the graph remains a trusted source of information.

5.3.5.1 Visualize and Explore Knowledge Graph

Once the knowledge graph has been appropriately constructed and deployed in a suitable storage system, the next step is to make it accessible to a range of user profiles, from business analysts to researchers and developers. To this end, interfaces and tools can be made available to perform queries, navigate between entities and relationships, and identify patterns or insights relevant to the business context.

Depending on the adopted technology, different query languages, such as SPARQL, Cypher, or GraphQL, may be supported. In parallel, it is common to implement interactive visualizations—such as dynamic graphs, dashboards, or reports—that facilitate the understanding of existing connections and the distribution of data in the domain. Adopting APIs or integration services also enables other corporate systems to consume the graph's information, expanding the project's organizational impact.

During this phase, business and IT teams observe user interaction with the graph and collect initial feedback on the usefulness of the features, ease of use, and query or analysis performance.

5.3.5.2 Evaluate Knowledge Graph

Systematically evaluating the KG quality, usefulness, and performance begins once the graph is used and its users can access the exploration features. The evaluation framework proposed here is divided into three layers:

1. **Intrinsic Data Quality:** Measures the syntactic and semantic correctness of the graph. Formal metrics such precision and recall are applied to extraction tasks. In contrast, consistency checks ensure that no A-Box instance violates T-Box constraints (using SHACL).

2. **Infrastructure Performance:** Evaluates the technological environment's capacity to handle the workload. It includes query response time (latency), throughput under concurrent access, and scalability of the triplestore/graph database.

3. **Business Utility:** Measures the graph's impact on decision-making. Through structured interviews and Likert-scale questionnaires, domain experts assess whether the KG provides insights that were previously unavailable or speeds up existing investigative processes.

By collecting feedback from end users — for example, through interviews, questionnaires, or monitoring usage in production systems — gaps, inconsistencies, and potential improvements are identified. It may lead to specific adjustments to the ontology, correcting imprecise definitions, updating classes and properties, or including new data sources that were not previously considered. These adjustments often lead to a return to previous stages, consolidating a continuous improvement cycle.

5.3.6 Support Activities

Whereas most of the process's activities are carried out at specific points, certain support activities continue throughout a project's lifecycle in conjunction with the primary actions in

knowledge graph development. These are not performed as separate tasks but rather accompany the rest of the stages and guide them so that the model remains conceptually correct, the data is handled appropriately, and everything occurs according to the organization's specific standards.

5.3.6.1 Ontology Improvement

This activity occurs continuously throughout the knowledge graph lifecycle, keeping the ontology up to date as discoveries or adjustments are identified. When exploring new data sources, extracting entities or relationships, and interacting with business stakeholders, concepts, properties, or restrictions not covered in the initial ontology may emerge. At this point, domain experts and knowledge engineers evaluate the relevance of these changes and analyze how they fit into existing definitions.

If approved, the new information is incorporated through an ontological refactoring process, with changes recorded in accordance with the established versioning rules. In addition, consistency and inference tests are performed to ensure that the revised ontology does not contain inconsistencies and remains aligned with business objectives.

Final validation, usually conducted with domain experts, ensures that the updates faithfully reflect the nuances of the organizational context. Upon completion, a revised, robust ontology is obtained, ready to guide subsequent data integration and analysis.

5.3.6.2 Data Governance

The Data Governance process also extends throughout the project execution, ensuring that the principles, policies, and controls defined in the planning are effectively applied. In this context, the governance team must monitor compliance with legislation, such as GDPR, and adopt measures to protect sensitive data, such as anonymization or access controls.

In addition, data quality is monitored at each stage of collection, pre-processing, extraction, and reconciliation, with metrics for completeness, consistency, and validity checked. When problems arise, the governance team guides corrections or forwards the need for revisions to the ontology through the Ontology Improvement step. A history of versioning and data lineage

is also maintained, recording when and by whom each entity or attribute was created, modified, or removed. Audits are conducted periodically, and compliance reports are generated to help ensure the knowledge graph maintains consistent adherence to institutional policies and defined quality standards.

5.4 LIMITATIONS AND RISKS

Despite its structured nature, the proposed process faces inherent methodological limitations. The first is the Expert Dependency risk; the process relies heavily on the availability and engagement of domain experts. In corporate settings, these profiles are often scarce, and their absence can stall the "Business Understanding"and "Evaluation"phases.

Technically, the Semantic Drift risk is another concern. As the graph progresses through the Knowledge Improvement stage, the gap between the original ontology and the ingested data may widen if the Ontology Improvement support activity is not strictly monitored. It can lead to inconsistencies in which the graph stores data that its schema no longer logically explains.

Lastly, there are performance risks related to Graph Bloat. In large-scale corporate environments, the overhead of maintaining high-granularity provenance and governance metadata for every triple can degrade query performance. Mitigation strategies, such as pruning old versions or using hybrid storage architectures, must be carefully planned in the Environment Planning phase.

5.5 SUMMARY

This chapter detailed the process for building Knowledge Graphs in corporate environments. This methodology overcomes the limitations of linear processes by formally integrating business understanding at the outset and establishing continuous data governance and ontology improvement as ongoing support activities. This approach ensures the graph's sustainability and relevance throughout its lifecycle.

Having now defined the two central methodologies—how to build the ontology and how to build the kg — the thesis moves from theory to practice. The following chapters will directly

apply these processes, using the methodologies to create the Asset Hiding Ontology, which will serve as the semantic foundation for the final implementation of the asset-hiding knowledge graph.

CHAPTER 6

PROPOSAL FOR AN ASSET CONCEALMENT ONTOLOGY

This chapter presents the execution of the ontology-building process to develop a domain ontology for asset concealment (TORRES *et al.*, 2024). The primary motivation for the ontology is to integrate multiple databases and structure the relationship among entities, assets, and financial transactions.

The ontology development begins with the application field and scope definition steps, which involve describing the techniques used to hide assets, mechanisms for investigating asset recovery, systematic methodologies for stakeholder identification, relevant data categories, and data integration challenges. The ontology will be built from multiple sources, including public records, government databases, financial transactions, and corporate registries, to cover all aspects of the domain.

Then, the ontology construction follows the steps defined in Chapter 4. The main entities, attributes, and relationships in the asset concealment area are also mapped in the specification activities. The implementation-related activities employ guidelines based on a systematic review of previous ontologies, aligning relevant concepts from the finance, legal, and relational domains to make them interoperable and extensible.

In the evaluation phase, the ontology data were validated and iteratively updated with real-world data in primary and secondary government databases to strengthen operational integration and completeness. It includes user validation, application testing, and data sampling ingestion and crossing, combined with fulfillment of the ontology for investigation and operational inquiry needs.

6.1 CONTEXTUALIZATION

The ontology for asset concealment investigation aims to validate the proposed methodology and address the challenge of integrating multiple databases, a requirement for investigative tools to uncover hidden assets. An ontology, being structured and providing a semantic representation of knowledge, enables data integration and improved understanding of the complex relationships among entities, assets, and financial transactions (TORRES *et al.*, 2024).

In addition to the latter, ontologies provide a means to identify hidden relationships, one of the significant hurdles to detecting asset concealment. Concealed assets are often spread out across complex networks of entities that appear significantly unrelated. An ontology helps define relationships such as "owns", "transfers to", or "is part of" and maps out elaborate schemes for transferring assets. This approach primarily exposes standard concealment methods that utilize shell companies, offshore tax havens, and fictive asset transfers—stratagems for concealing true ownership.

Another significant advantage of using ontologies is that they can integrate diverse data sources. Information on people's income, expenditures, and financial flows comes from sources such as public records, banks, court filings, social media, and even government records. An ontology will help organize these disparate datasets so that the data can be systematically linked and analyzed, breaking down information silos and enabling more rigorous investigations.

By using an ontological structure for their data, AI systems can also identify potentially suspicious asset movements more efficiently. Once this ontological structure has been developed, machine learning algorithms can detect patterns in financial transactions that suggest a specific attempt at concealment. For instance, AI-powered systems can automatically flag assets that have been transferred back and forth among a group of interconnected entities—a key sign of money laundering or attempts to conceal assets (TORRES *et al.*, 2024).

6.2 SPECIFICATION

The initial phase of the process was carried out through a collaborative effort between Knowledge Engineers and domain specialists in asset concealment fraud. This fact happened

within the scope of a cooperation term between the University of Brasilia and the Attorney General's Office of Brazil. The main aim is to gain an in-depth understanding of what really happens in the investigative process of fraud detection, to accurately specify the project's scope, to systematize the capture of domain knowledge, and to evaluate whether the existing related ontologies are applicable (TORRES *et al.*, 2024).

6.2.1 Application Field and Scope Definition

Asset concealment encompasses different aspects of asset identification, tracing, and recovery. Those who thwart justice might resort to evasive tactics to avoid liability for tax obligations, debts arising from creditor claims, or regulatory scrutiny. The criminals use various methods to conceal their ownership, control, or existence of the assets, including complex incorporation arrangements, offshore banking, shell companies, and falsified or manipulated records to obscure asset tracing.

On the other hand, recovery refers to the investigative or legal efforts needed to locate and recover an asset that has been hidden or acquired illegally. Most asset recovery efforts involve legal disputes, financial fraud, or corruption investigations. Usually based on forensic accounting, litigation, international collaboration, and other technological tools to trace financial debts, recoveries typically rely on investigating hidden ownership networks. The objective of such activities is to address repatriation to the government, the creditor, or the member.

During discussions with domain experts, a number of behaviors associated with asset concealment fraud were identified. With a brief description of these behaviors, this list is presented in Table 6.1 (TORRES *et al.*, 2024). Reviewing the unique nature of these persons, specific to fraudulent concealment, also enhanced the analysis of the major actors involved and the standard fraudulent methods used to conceal assets. Emphasis was placed on tracing the relationships among the persons, businesses, and assets. Also, strong attention was paid to government inputs since most asset concealment is directly related to tax evasion. Therefore, the ontology must describe the concepts relevant to this specific type of fraud across several jurisdictions.

Table 6.1: Behaviors associated with Asset Concealment fraud. (TORRES *et al.*, 2024).

Id	Description
1	People who commit asset concealment fraud often buy high-value items such as automobiles, boats, aircraft, or properties to convert liquid assets into physical ones that are harder to trace.
2	Fraudsters often place their assets in the name of relatives, such as parents, siblings, or spouses, to mask the true owner and avoid asset seizure.
3	People also tend to conceal assets by holding them with friends, neighbors, or close associates, further distancing the true owner from the property.
4	Individuals hide assets by creating or using existing companies, often setting up complex ownership structures with distributed shareholding to obscure the true ownership. This can involve shell companies or multiple layers of corporate entities across different jurisdictions.
5	Fraudsters often transfer assets to offshore bank accounts in countries with strict privacy laws, making it difficult for authorities to trace the funds.
6	Increasingly, people use cryptocurrencies to hide assets, as these can be more challenging to track due to their decentralized nature and the anonymity they can provide.
7	Governments are among the primary victims of asset concealment fraud despite commonly having databases related to individuals, their attributes, and assets.
8	Governments commonly lack a strategy for integrating distributed databases related to the same entities. In addition, similar data stored in different databases often exhibit inconsistencies.

When it comes to the data needed to analyze the behaviors listed in Table 6.1, it is clear that much of these data are held by government agencies. The problem is that these data are scattered across different government agencies or even governments in other countries, given the transnational character of asset concealment crimes. Relevant datasets are identified by selecting increasingly informative databases of individuals and businesses, covering a vast array of tangible and intangible assets and including attributes that facilitate connections among those entities. These obstacles underscore the need to develop this ontology to support data integration and analysis.

6.2.2 Data Cluster Identification

According to the information identified in the previous step, the domain of asset concealment analysis is closely linked to people, companies, assets, and relationships that directly or indirectly link them (TORRES *et al.*, 2024). We defined six data categories to guide this search

process, based on internal databases provided by the Brazilian public agency. Most of these categories contain data held by governments, since this actor is also the one with the most data about people and their possessions, and is often a primary victim of this type of fraud. This strategy facilitates the identification of related data, avoiding the collection of data outside the scope. For this reason, the definition of categories must be applied regardless of the domain under analysis.

Table 6.2: Categories of datasets related to asset concealment.(TORRES *et al.*, 2024).

Category	**Description**
Biographical data of People	Information on births and deaths, marriage and Divorce, Citizen Identification Registry and Passports, Voters or Licensed Drivers Database.
Healthcare and Educational Data	Information on healthcare plans, educational records of people enrolled in universities, data of members of research groups, and data from scientific papers databases.
Taxes, Finances, Labor and Employment	Information on tax declarations, assets, and payments of citizens; Information on social security contributions and benefits; Employment Records - Employment history, data on employers, and positions held.
Data from Companies	Registration information includes company name, registration number, date of incorporation, ownership details, financial records, etc.
Judicial Processes Data	Case Information includes the case number, type of case (civil, criminal, administrative), court jurisdiction, and the parties involved, such as plaintiffs, defendants, lawyers, and judges.
Data from Assets and Owners	Motor, Watercraft, and Aircraft Vehicle Registration Database, Real State Owners Database, Asset owner information for public servants, Information on bank accounts, Stock purchase and government bond information, Income tax declaration, Patent, Trademark, and Copyright public search databases.

[1] Some of the data mapped can be protected by confidentiality, even among government agencies.

Based on the information developed in the previous step, the domain of asset concealment analysis is closely linked to persons, companies, assets, and the relationships that connect

them directly or indirectly. Six data categories (Table 6.2) were defined to guide this search procedure, and they were also derived from internal databases supplied by the Brazilian public agency (UNIVERSITY OF BRASILIA, 2025). Most of these categories involve data generated by governments, for governments are the actors with the most information about people and their property, and they are often among the leading actors affected by this kind of fraud (TORRES *et al.*, 2024). This strategy facilitates the identification of associated data simply, thus avoiding the selection of data outside the set guidelines. For this reason, the definition of categories should be applied regardless of the domain being analyzed.

6.2.3 Domain Knowledge Capture

In the preliminary stages, metadata from the identified categories was examined to gain knowledge of the domain. It is also essential to determine a set of relevant actors for analysis. Accordingly, Brazil, Canada, the United States of America (USA), France, Germany, the United Kingdom (UK), India, and China were selected based on economic relevance, regional importance, and the availability of open, searchable metadata from government databases. The domain-capture process involved mapping the countries involved for these identified categories (Table 6.2). The main emphasis is on identifying core classes, attributes, and relationships relevant to the specific domain.

This work presents the study by (TORRES *et al.*, 2024) in the following subsections, where each category is analyzed in the context of the selected countries, and the collection and processing of data for each category are discussed in detail.

6.2.3.1 Biographical data of People

The records of births, deaths, marriages, and divorces are often managed by civil registries, which are the official repositories for national government administrations. Other primary databases included in this category are passports, voter registration, and driver's licenses. These databases provide biographical and demographic data across several areas, such as legal documentation, policy formulation, and resource allocation, and include identifiers for citizens, such as national ID numbers, to establish citizen profiles with specific biographical data.

The responsibility for managing the various databases related to citizens' biographical data is often distributed across multiple institutions in countries. The analysis presented in (TORRES *et al.*, 2024) highlights the primary related databases and attributes, as well as the bodies responsible for managing these databases in each country. About Civil Registration Systems/-Databases, the following relevant databases and attributes were identified:

- **Systems or Databases**—Civil Registry Offices—Brazil, Person General Register - RG (varies by state—Brazil), Service Public d'État Civil Databases (Public Civil Status Service—France), Standesamt Databases (Civil Registry Office—Germany), General Register Office Database (GRO—UK), Hukou System (China), Aadhaar (India).
- **Common Attributes Present**—Unique ID, name, date of birth, place of birth, parents' names, birthplace, date of death, place of death, names of spouses, date of marriage, place of marriage, address of residence, contact information (email or phone number).

(TORRES *et al.*, 2024) also identified the following relevant systems and attributes related to Electoral Registration Systems or Databases.

- **Systems or Databases**—Voter Registration Database—TSE Database—Brazil, National Register of Electors (Elections Canada), Répertoire Électoral Unique (REU) (National Voter Database—France), Wählerverzeichnis (Voter List—Germany), Electoral Registers (Managed by Local Authorities—UK), Electoral Roll (Election Commission of India).
- **Common Attributes Present**—Voter ID number, name, date of birth, address, polling station, voting district, eligibility status, contact information (email or phone number), biometric data (for example, in Brazil and India).

A similar survey was conducted with Passport Control Systems/Databases (TORRES *et al.*, 2024):

- **Systems or Databases**—Passport System (Federal Police—Brazil), Passport Information Electronic Records System (PIERS) (Department of State—US), Canadian Passport Program (Immigration, Refugees and Citizenship Canada—IRCC), Système de Gestion

des Passeports (Ministry of the Interior—France), Passregister (Federal Office of Administration—Germany), Passport Application and Issuance System (His Majesty's Passport Office (UK), National Immigration Administration Passport Database (China), Passport Seva (Ministry of External Affairs—India).

- **Common Attributes Present**—Passport number, name, date of birth, nationality, date of issue, expiration date, place of issue, biometric data.

Finally, at the end of this category, attributes and systems related to driver registration data were identified (TORRES *et al.*, 2024):

- **Systems or Databases**—Brazilian National Driver's License Database (RENACH—Brazil), Department of Motor Vehicles or equivalent agencies databases (varies by state—US), Fichier National des Permis de Conduire (National Driver's License File—France), Fahrerlaubnisregister (Federal Motor Transport Authority—Germany), DVLA Driver Database (Driver and Vehicle Licensing Agency—UK), Driver License System (Ministry of Public Security—China), Sarathi (National Register of Driving Licenses—India).

- **Common Attributes Present**—Driver's license number, name, date of birth, category of license, parents' names, status of license (active, suspended), date of issue, address, expiration date.

As it was presented, civil registries, voters, driver's licenses, and passport databases are crucial across various countries as they store essential information about citizens, including biographical data that individualizes a person, such as a name, parents' names, birth date, home address, email, and phone number (TORRES *et al.*, 2024). At the same time, these attributes can link people who share the same mother or father, live in the same home, or are neighbors, for example. The relevant entities, attributes, and relationships identified in these categories based on the data analyzed are presented in Table 6.3 (TORRES *et al.*, 2024).

Table 6.3: Entities, Attributes, and Relationships identified in Biographical data of People. (TORRES *et al.*, 2024)

Attributes	**Entities and Relationships**
• Person: unique_id, name, name_of_mother, name_of_father, date_of_birth, birthplace • Email: address. • PhoneNumber: country_code, local_code, number. • Address: street, number, complement. • ZipCode: number, city, state, country.	$Person \xleftarrow[IS\_MARRIED]{} Person$ $Person \xleftarrow[IS\_DIVORCED]{} Person$ $Person \xrightarrow{IS\_FATHER} Person$ $Person \xrightarrow{IS\_MOTHER} Person$ $Person \xrightarrow{HAS} PhoneNumber$ $Person \xrightarrow{HAS} Email$ $Person \xrightarrow{HAS} Address$ $Address \xrightarrow{HASR} ZipCode$

6.2.3.2 Healthcare and Educational Data

This data category encompasses a wide range of information, from patient medical records to student information. Although many governments use this data to manage public health and educational policies and enhance the overall health and education of the population, some databases, such as medical records, are protected by confidentiality laws. About Healthcare Data Systems/Databases, the following relevant databases and attributes were identified (TORRES *et al.*, 2024):

- **Systems or Databases**—National Health Card Registration (CADSUS—Brazil), Medicare and Medicaid (US), Système National des Données de Santé (SNDS—France), NHS Spine (UK), Ayushman Bharat Databases and National Health Stack (India).
- **Common Attributes Present**—Unique ID, patient ID, name, parent's name, spouse's name, date of birth, medical history, treatments, medications, hospital visits, diagnostic results, vaccination status, address, contact information (ex. email or phone number).

A similar survey was conducted with Educational Systems/Databases (TORRES *et al.*,

2024):

- **Systems or Databases**—Educacenso System and Lattes Platform (Brazil), National Center for Education Statistics Databases (US), Canadian Centre for Education Statistics Databases (Canada), Système d'Information sur le Suivi de l'Élève (SISE—France), German Academic Exchange Service and Destatis Databases (Germany), National Pupil Database (NPD—UK), Center for Student Services and Development Databases (China), Unified District Information System for Education Plus (UDISE±–India).
- **Common Attributes Present**—Student ID, name, date of birth, university or school enrollment, grade level, attendance records, academic performance, graduation status, research groups with members' names (Lattes Platform).

Table 6.4: Healthcare and Educational Data. (TORRES *et al.*, 2024)

Attributes	**Entities and Relationships**
• Person: unique_id, name, name_of_mother, name_of_father, date_of_birth, birthplace, name_of_children, name_of_spouse • Paper: doi, title, authors, authors_affiliation, authors_emails • ResearchGroup: title, id, members • Class: name, university, class_id, students • Company (University): name, company_id, classes • Email: address. • PhoneNumber: country_code, local_code, number. • Address: street, number, complement. • ZipCode: number, city, state, country.	$Person \xrightarrow{IS\_MEMBER\_OF} ResearchGroup$ $Person \xrightarrow{IS\_AUTHOR\_OF} Paper$ $Person \xrightarrow{ATTEND} Class$ $Company \xrightarrow{OFFER} Class$ $Person \xrightarrow{IS\_FATHER} Person$ $Person \xrightarrow{IS\_MOTHER} Person$ $Person \xrightarrow{HAS} PhoneNumber$ $Person \xrightarrow{HAS} Email$ $Person \xrightarrow{HAS} Address$ $Address \xrightarrow{HAS} ZipCode$ $Person \xrightarrow{IS\_CHILD} Person$ $Person \xleftarrow[IS\_MARRIED]{} Person$

Due to confidentiality issues, full access to the bases was not always possible in this category. However, the public description of attributes in most of the presented databases and metadata

collected from the websites of government agencies of the countries analyzed made it possible to extract the entities, attributes, and relationships presented in Table 6.4 (TORRES *et al.*, 2024).

6.2.3.3 Taxes, Finances, Labor and Employment

These data are among the most essential managed by governments. Payment of fees and taxes is a common source of income for countries. In addition, income tax declarations and employee databases greatly assist governments in effectively collecting these fees and taxes. By the way, social services have two aspects: social security in retirement, and, second, more common in poorer countries, helping people with lower incomes through income distribution policies. About Taxes and Finances Data Systems/Databases, the following relevant databases and attributes were identified (TORRES *et al.*, 2024):

- **Systems or Databases**—Individual Taxpayer Registry - Brazil, Internal Revenue Service Database (IRS—US), Canada Revenue Agency Tax System (Canada), Système Fiscal de la Direction Générale des Finances Publiques (DGFiP—France), His Majesty's Revenue and Customs Databases, State Taxation Administration Tax System (China), Income Tax Department e-Filing Portal (India).
- **Common Attributes Present**—Unique ID, taxpayer ID, name, income, deductions, tax payments, tax returns, employment details, assets ownership data, contact information (email or phone number), address.

Finally, at the end of this category, attributes and systems related to Labor and Employment Systems/Databases were identified (TORRES *et al.*, 2024):

- **Systems or Databases**—General Register of Employed and Unemployed Persons in Brazil (CAGED—Brazil), Annual Social Information Report of Brazil (RAIS—Brazil), Canada Record of Employment System, Déclaration Sociale Nominative (France), Bundesagentur für Arbeit Database (Federal Employment Agency—Germany), Employee's Provident Fund Organisation Databases (India), Ministry of Human Resources and Social Security Database (China).

- **Common Attributes Present**—Unique ID, employee ID, name, parent's name, spouse's name, birthplace, employment status, job title, salary, employer, start and end dates, social security contributions, address.

Table 6.5: Taxes, Finances, Labor and Employment. (TORRES *et al.*, 2024)

Attributes	**Entities and Relationships**
• Person: unique_id, name, name_of_mother, name_of_father, date_of_birth, birthplace, name_of_children, name_of_spouse • Company: name, company_id, employees • Asset/BankAccount: type, description, id, owner, monetary_value • Email: address. • PhoneNumber: country_code, local_code, number. • Address: street, number, complement. • ZipCode: number, city, state, country.	$Person \xrightarrow{IS\_CHILD} Person$ $Person \xleftrightarrow{IS\_MARRIED} Person$ $Person \xrightarrow{HAS} PhoneNumber$ $Person \xrightarrow{HAS} Email$ $Person \xrightarrow{HAS} Address$ $Address \xrightarrow{HAS} ZipCode$ $People \xrightarrow{WORKS\_IN} Company$ $People \xrightarrow{HAS} Asset/BankAccount$ $People \xrightarrow{HAS} Asset/CryptoWallet$ $People \xrightarrow{HAS} Asset/Salary$ $Person \xrightarrow{MADE\_WIRE\_TRANSFER} Person$ $CryptoWallet \xrightarrow{TRANSFER} CryptoWallet$ $Person \xrightarrow{IS\_FATHER} Person$ $Person \xrightarrow{IS\_MOTHER} Person$ $Person \xleftrightarrow{IS\_FAMILY} Person$

As presented, tax and banking data, as well as income tax declarations, are generally protected by confidentiality and are accessible only to the agencies directly responsible for storing them. Due to their scope, these databases are among the most important because they allow the definition of relationships between people and assets (TORRES *et al.*, 2024). In any case, confidentiality can be breached if the state charges a person. It is also important to highlight that a large part of the assets listed in the income tax declaration is registered in the original databases of the bodies that control their ownership (TORRES *et al.*, 2024).

The register of employed people not only allows us to have an idea of people's income but also helps to identify co-workers' relationships. Finally, social service data also helps, among other things, to determine income and, sometimes, to identify family members related to beneficiaries (TORRES *et al.*, 2024).

The entities, attributes, and relationships identified in this category of data are shown in Table 6.5 (TORRES *et al.*, 2024).

6.2.3.4 Judicial Processes Data

Many countries maintain databases to manage judicial processes and legal cases. These systems are essential for ensuring the rule of law, facilitating access to justice, and maintaining transparency and accountability in the legal system. These databases include detailed information on ongoing and concluded cases, such as case numbers, parties involved, legal representatives, case status, court decisions, and hearing schedules. It also tracks case filings, motions, and judicial rulings to ensure transparency and accessibility for all parties involved (TORRES *et al.*, 2024). About Judicial Processes Systems/Databases, the following relevant databases and attributes were identified (TORRES *et al.*, 2024). The elements presented in Table 6.6 were mapped using these attributes.

- **Systems or Databases**—Electronic Judicial Process (PJe), Public Access to Court Electronic Records (PACER—US), France Portalis (Ministère de la Justice—France), Germany Elektronischer Rechtsverkehr, His Majesty's Courts and Tribunals Service (United Kingdom).
- **Common Attributes Present**—Case number, parties involved, type of case, court, status, filings, decisions, hearing dates, motions, judgments, appeals, legal representation.

6.2.3.5 Data from Companies

Government databases that maintain data about local companies are relevant in supporting economic development, regulatory compliance, and public transparency (TORRES *et al.*, 2024). These databases collect and store extensive information about businesses operating within a

jurisdiction, providing valuable resources for government agencies, businesses, and the public (TORRES *et al.*, 2024). Regarding data from Companies' Systems/Databases, the following relevant databases and attributes were identified (TORRES *et al.*, 2024).

- **Systems or Databases**—National Registry of Legal Entities (CNPJ—Brazil), Corporations Canada Database, EDGAR database (US), Registre National des Entreprises (France), Companies House (UK), Ministry of Corporate Affairs—Master Data Services V3 (India).

- **Common Attributes Present**—Company ID, company name, trade name, date of incorporation, business activity, address, legal representative, directors, status (active/inactive).

Table 6.6: Judicial Processes Data. (TORRES *et al.*, 2024)

Attributes	**Entities and Relationships**
• Person: unique_id, name • Company: name, company_id • LawSuite: unique_id, type, court, lawyers, plaintiffs, defendants, judge, monetary_value. • Asset: type, description, id, owner, monetary_value • Email: address. • PhoneNumber: country_code, local_code, number. • Address: street, number, complement. • ZipCode: number, city, state, country.	$Person \xrightarrow{IS\_LAWIER} LawSuite$ $Person \xrightarrow{IS\_DEFENDANT} LawSuite$ $Person \xrightarrow{IS\_PLAINTIFF} LawSuite$ $Company \xrightarrow{IS\_PLAINTIFF} LawSuite$ $Company \xrightarrow{IS\_DEFENDANT} LawSuite$ $Company \xrightarrow{IS\_JUDGE} LawSuite$ $LawSuite \xrightarrow{HAS} Asset$ $Person \xrightarrow{HAS} PhoneNumber$ $Person \xrightarrow{HAS} Email$ $Person \xrightarrow{HAS} Address$ $Address \xrightarrow{HAS} ZipCode$ $Company \xrightarrow{HAS} PhoneNumber$ $Company \xrightarrow{HAS} Email$ $Company \xrightarrow{HAS} Address$

Table 6.7: Data from Companies. (TORRES *et al.*, 2024)

Attributes	**Entities and Relationships**
• Person: unique_id, name • Company: name, company_id, monetary_value, partners • Asset: type, description, id, owner, monetary_value • Email: address. • PhoneNumber: country_code, local_code, number. • Address: street, number, complement. • ZipCode: number, city, state, country.	$Person \xrightarrow{IS\_PARTNER} Company$ $Company \xrightarrow{IS\_PARTNER} Company$ $Company \xrightarrow{HAS} Asset$ $Company \xrightarrow{HAS} PhoneNumber$ $Company \xrightarrow{HAS} Email$ $Company \xrightarrow{HAS} Address$ $Address \xrightarrow{HAS} ZipCode$

Among the data commonly found in this type of database, it is important to highlight information on a company's partners and the company's monetary value. This information is crucial, as establishing a company is one of the primary methods for concealing assets. The entities and relationships presented in Table 6.7 were mapped using the attributes found in these databases.(TORRES *et al.*, 2024)

6.2.3.6 Data from Assets and Owners

Data on assets associated with individuals is among the most relevant information for governments, especially in combating crime (TORRES *et al.*, 2024). For this reason, there are extensive databases associated with this purpose. It is essential to note, however, that much of the data in this category of databases is subject to varying levels of confidentiality and can be accessed only with judicial authorization. The following list describes the databases and attributes identified in searching only open source data, related to the category Data from Assets (TORRES *et al.*, 2024):

- **Systems or Databases**—RENAVAM (National Registry of Motor Vehicles—Brazil), Department of Motor Vehicles—by US state, Système d'Immatriculation des Véhicules

(Vehicle Registration System—France), KBA (Federal Motor Transport Authority—Germany), Driver and Vehicle Licensing Agency—Vehicle Records (UK), Vahan (National Register of Motor Vehicles—India), Vessel Registration (Brazilian Navy—Brazil), Canadian Register of Vessels (Transport Canada), USCG National Vessel Documentation Center (US Coast Guard—US), Registre International Français (France), German International Shipping Register (Germany), UK Ship Register (Maritime and Coastguard Agency—UK), Brazilian Aeronautical Registry (ANAC—Brazilian Civil Aviation Agency), Canadian Civil Aircraft Register Computer System (CCARCS—Canada), FAA Aircraft Registry (Federal Aviation Administration—US), Registre des Aéronefs Civils (French Civil Aviation Authority—France), Luftfahrt-Bundesamt Database (Germany), G-INFO (Civil Aviation Authority—UK), DGCA Aircraft Register (Directorate General of Civil Aviation—India), Rural Property Registry (CAFIR—Brazil), Environmental Rural Register (CAR—Brazil), Certificate of Registration of Rural Property (CCIR—Brazil), Ontario Land Registry Access (Canada), HM Land Registry Database—UK, Registrato (Brazil), Declaration of Assets of Public Servants (Brazil), Declaration of Assets of Candidates in Electoral Dispute (Brazil), United States Patent and Trademark Office Database (USPTO), European Patent Office Database (EPO), European Union Intellectual Property Office Database, United States Copyright Office Public Catalog.

- **Common Attributes Present**—Aircraft: registration/serial number, aircraft type, manufacturer, year of manufacture. Vessel: registration number, hull identification, owner, type of vessel, tonnage, year of construction. Motor Vehicles: license plate number, chassis number, owner, vehicle type, year of manufacture, category. Real Estate: property ID, owner name, location/address, property type (urban/rural), size, cadastral number, legal description, date of acquisition, zoning information, boundaries, liens or encumbrances, and how much the property was last sold for. Assets in General: monetary value, type, description.

Using these databases and attributes, it was possible to identify a set of assets related to specific types (Table 6.8) (TORRES *et al.*, 2024). Although it is not, once again, an exhaustive list, it allows the identification of a set of assets that can significantly assist in searching for hidden assets.

Table 6.8: Data from Assets and Owners. (TORRES *et al.*, 2024)

Attributes	**Entities and Relationships**
• Person: unique_id, name • Company: name, company_id • Asset: type, description, id, owner, monetary_value	$Person/Company \xrightarrow{HAS} Asset : MotorVehicle$ $Person/Company \xrightarrow{HAS} Asset : Watercraft)$ $Person/Company \xrightarrow{HAS} Asset : Aircraft$ $Person/Company \xrightarrow{HAS} Asset : House$ $Person/Company \xrightarrow{HAS} Asset : Apartment$ $Person/Company \xrightarrow{HAS} Asset : Land$ $Person/Company \xrightarrow{HAS} Asset : Building$ $Person/Company \xrightarrow{HAS} Asset : BankAccount$ $Person/Company \xrightarrow{HAS} Asset : GovBound$ $Person/Company \xrightarrow{HAS} Asset : StockShares$ $Person/Company \xrightarrow{HAS} Asset : Patent$ $Person/Company \xrightarrow{HAS} Asset : Trademark$ $Person/Company \xrightarrow{HAS} Asset : Copyright$

6.2.4 Domain Knowledge Documentation

Throughout the process, concepts related to the domain under analysis were identified and documented. At the end of domain knowledge capture, these concepts identified by category are combined into a single structure, presented in Figure 6.1 (TORRES *et al.*, 2024).

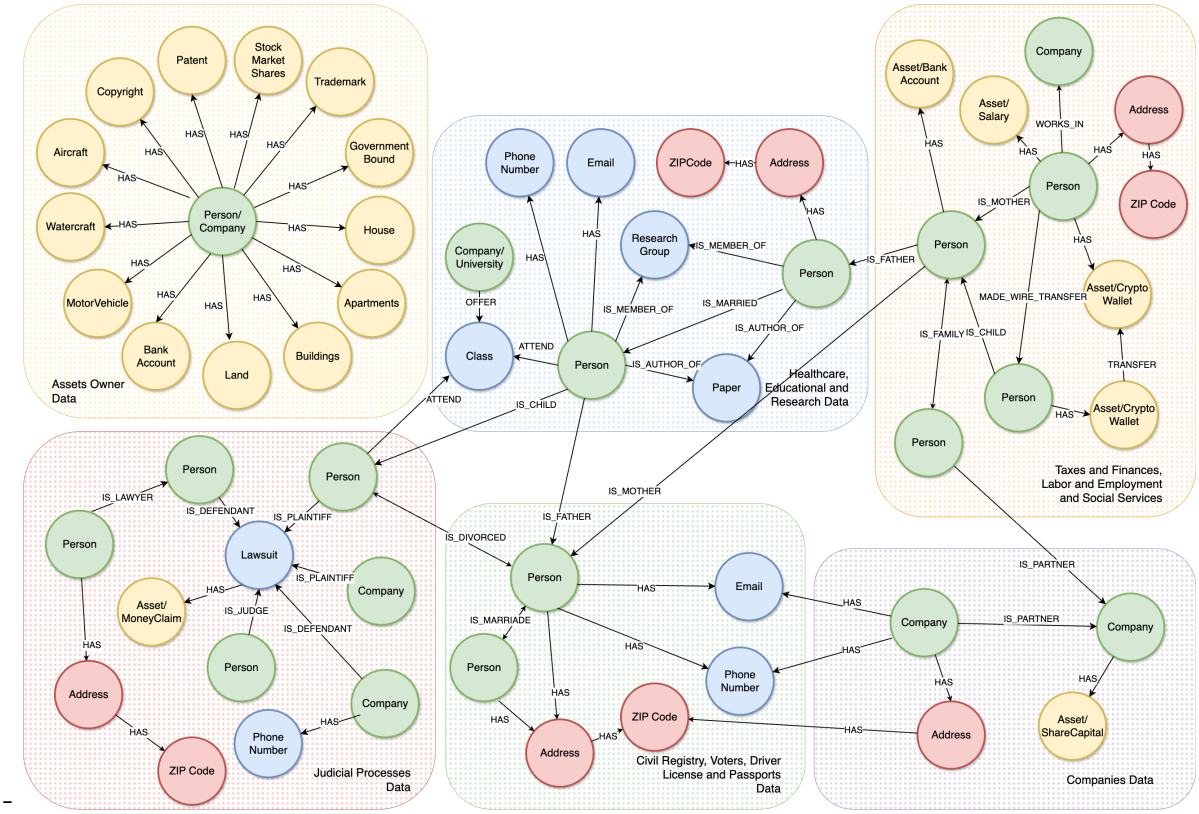

Figure 6.1: Asset Concealment Preliminary Ontology. (TORRES *et al.*, 2024)

6.2.5 Existing Ontology Reusability Review

Asset concealment is a specific field concerned with identifying individuals, their assets, and the strengths or weaknesses that facilitate their concealment. So far, we are not aware of any existing ontology or knowledge graph that covers this domain. However, as already presented, several concepts and relationships related to asset concealment overlap with other domains, particularly in finance and personal relationships, including familial and social ties.

In this sense, some related ontologies were used as input. Torres (TORRES *et al.*, 2024) selected the Financial Industry Business Ontology (BENNETT, 2013), the Suggested Upper Merged Ontology (NILES; PEASE, 2001) with its specific domain ontology SUMO Financial, and the ONTO-FIC ontology (ABROUK *et al.*, 2023) from the Financial Domain. Ontouml (CARVALHO *et al.*, 2014), the Family History Knowledge Base (STEVENS *et al.*, 2014), the Kinship Ontology (CHUI *et al.*, 2020), and the Friend of a Friend ontology (BRICKLEY; MILLER, 2014) were chosen from the Kinship and Friendship domain (TORRES *et al.*, 2024).

6.3 IMPLEMENTATION

The implementation phase involves three closely interconnected and iterative tasks: defining ontology classes, relationships, and attributes. These elements are developed in an integrated manner, as their identification processes are interdependent. During this stage, the knowledge gathered from stakeholders and the information extracted from relevant database categories serve as inputs. Additionally, the ontologies identified during the review of existing ontology reusability must be aligned to ensure consistency and interoperability.

6.3.1 Existing Ontology Alignment

The main concepts related to the domain in analysis were captured from FIBO (BENNETT, 2013), SUMO-F (NILES; PEASE, 2001), FHKB (STEVENS *et al.*, 2014), T-Kinship (CHUI *et al.*, 2020), FOAF (BRICKLEY; MILLER, 2014), OntoUML (CARVALHO *et al.*, 2014), and Onto-FIC (ABROUK *et al.*, 2023), and aligned (TORRES *et al.*, 2024). Subsequently, a correlation was established between similar concepts with different identifiers, as presented in Table 6.9.

Table 6.9: Financial Ontology Alignment (TORRES *et al.*, 2024)

Concept Description	**FIBO**	**SUMO**	**FHKB**	**T-Kinship**	**FOAF**	**Onto-FIC**
A corporate or similar institution.	Organization	Organization	Company	-	-	Organization
Modern man, the only remaining species of the human genus	Person	Human	Person	Person	Person	Person
A Collection of Agents (Person or Company)	Group	Group	-	-	Group	-
Physical address where representatives located for any kind of organization or person	Physical Address	Physical Address	-	-	-	-

Continued on next page

Table 6.9 – continued from previous page

Concept Description	FIBO	SUMO	FHKB	T-Kinship	FOAF	OntoFIC
Address identifying a virtual, i.e. non-physical, location	-	Virtual Address	Email Address	-	-	-
t is a phone number corresponding to the Telephone	-	Phone Number	-	-	-	-
Any item of economic value	Good	Financial Asset	-	-	-	-
Relatively permanent enclosed structure, that has a roof and walls and stands more or less permanently in one place	Building	Building	-	-	-	-
A Certificate that expresses the content of an invention that has been accorded legal protection by a governmental entity	-	Patent	-	-	-	-
Where something with economic value is exchanged for something else	Transaction Event	Financial Transaction	Transaction	-	-	-
An Activity of money being transferred into a customer's account at a financial institution	-	Deposit	-	-	-	-
An activity of money being transferred from a customer's account at a financial institution	-	WithDrawal	-	-	-	-
An activity of committing money or capital in order to gain a financial return	-	Investing	-	-	-	-

Continued on next page

Table 6.9 – continued from previous page

Concept Description	FIBO	SUMO	FHKB	T-Kinship	FOAF	Onto-FIC
Delivery of money in fulfillment of an obligation, such as to satisfy a claim or debt	Payment	Payment	-	-	-	-
The chain of events of selling all of a company's assets, paying outstanding debts, and distribution of the remainder to shareholders, and them going out of business.	Liquidation	Liquidation	-	-	-	-
Any transaction which involves Stock and which occurs in a Stock Market.	-	Stock Market Transaction	-	-	-	-
An activity of opening a financial account	-	Opening An Account	-	-	-	-
An item of value purchased for income or capital appreciation.	-	Investment	-	-	-	-
A debt instrument issued for a period of more than one year with the purpose of raising capital by borrowing.	Bond	Bond	-	-	-	-
An instrument that signifies an ownership position, or equity, in a Corporation, and represents a claim on its proportionate share in the corporation's assets and profits	-	Stock	-	-	-	-

Continued on next page

Table 6.9 – continued from previous page

Concept Description	FIBO	SUMO	FHKB	T-Kinship	FOAF	Onto-FIC
Land, including all the natural resources and permanent buildings on it.	Real Estate	Real Estate	-	-	-	-
A relatively short item that either is unbound or is bound with other Articles in a Book or a Scientific Paper	-	Article	-	-	-	-

Beyond aligning concepts, an analysis of related properties was conducted to map relationships, focusing on connections among individuals, especially within families and social networks. As illustrated in Table 6.10, most reviewed ontologies included only a limited set of relationships for constructing family genealogies. Moreover, these are traditionally centered upon connecting parent-child and spousal relationships from which further family ties could be inferred. As this study shows, however, it is crucial to explicitly define each tie, since the parent-child link may not always exist or be relevant.

As far as the situation of just one sibling is concerned, the case of an individual without parental information is essential. In such a case, one can characterize this relationship by a specific property rather than inferring it from parent-child relationships. This property would be functionally synonymous with ascribing to two individuals who share at least one set of parents. This type of ambiguity manifests as "grandfather" relationships, represented as a forward chain of father/child spans, in which a person is the father of the father of the father of the mother.

Table 6.10: Genealogy Ontology Alignment (TORRES *et al.*, 2024)

Feature	FHKB	ONTO-UML	T-KINSHIP
Parentage	isParentOf	-	hasChild/hasParent
	isParentOf >isFatherOf	MFatherOf	-
	isParentOf >isMotherOf	MMotherOf	-
	has_relation >isAncestorOf	MAncestorOf	ancestorOf

Continued on next page

Table 6.10 – continued from previous page

Feature	FHKB	ONTO-UML	T-KINSHIP
Grandparents and Great Grandparents	isGrandParentOf	-	-
	isGrandmotherOf	-	-
	isGrandFatherOf	-	-
Aunts and Uncles	has_relation >isBloodrelationOf >isUncleOf	-	-
	has_relation >isBloodrelationOf >isAuntOf	-	-
	has_relation >isBloodrelationOf >isGreatUncleOf	-	-
	has_relation >isBloodrelationOf >isGreatAuntOf	-	-
Siblings	has_relation >isBloodrelationOf >isSiblingOf	-	-
	has_relation >isBloodrelationOf >isSiblingOf >isBrotherOf	-	-
	has_relation >isBloodrelationOf >isSiblingOf >isSisterOf	-	-
Cousins	isBloodRelationOf-IsCousinOf	-	-
	isBloodRelationOf >isCousinOf >isFirstCousinOf	-	-
	isBloodRelationOf >isCousinOf >isSecondCousinOf	-	-
	isBloodRelationOf >isCousinOf >isFirstCousinOnceRemovedOf	-	-
Marriage	isPartnerIn	-	-
	isSpouseOf	-	hasSpouse
	isSpouseOf >isWifeOf	-	-
	isSpouseOf >isHusbandOf	-	-
InLaw	isInLawOf >isParentInLawOf	-	-
	isInLawOf >isParentInLawOf >isMotherInLawOf	-	-
	isInLawOf >isParentInLawOf >isFatherInLawOf	-	-
	isInLawOf >isSiblingInLawOf	-	-
	isInLawOf >isSiblingInLawOf >isSisterInLawOf	-	-
	isInLawOf >isSiblingInLawOf >isBrotherInLaw	-	-
	isInLawOf >isAuntInLawOf	-	-
	isInLawOf >isUncleInLawOf	-	-

6.3.2 Ontology Construction

This ontology development process fused conceptions adopted from current ontologies with those nominated from domain analysis. The figure 6.2 illustrates the final collection of selected concepts and their imposing structure in the asset concealment domain (TORRES *et al.*, 2024). The taxonomy is to promote the effectiveness and accuracy of investigations by providing a strategic framework for identifying, analyzing, and tracing hidden assets.

Additionally, the object properties of the ontology were defined (Figure 6.3). Object properties define relationships between entities or concepts within a domain in ontology development (TORRES *et al.*, 2024).

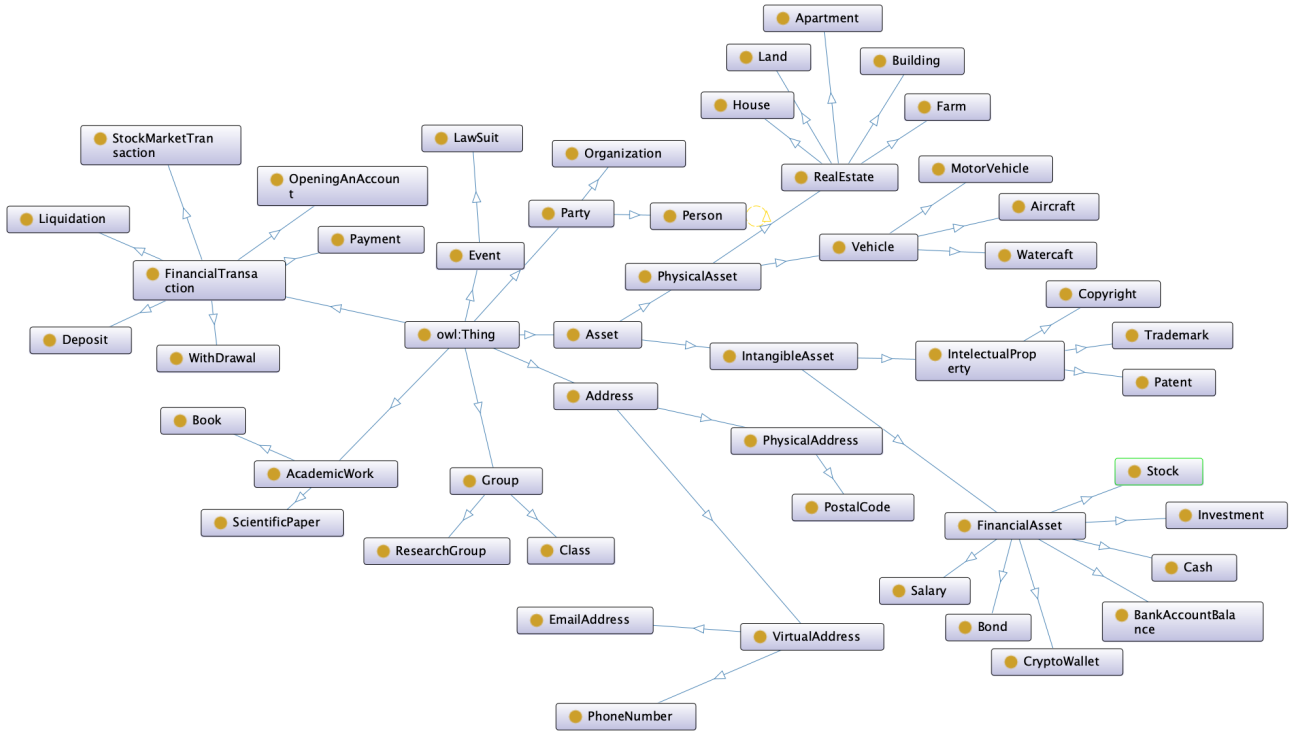

Figure 6.2: Asset Concealment Concepts (TORRES *et al.*, 2024)

6.4 ONTOLOGY EVALUATION

The ontology generated for asset hiding was evaluated using a three-step validation. This method bases the evaluation on the user and on the application of the ontology, following the proposed process (TORRES *et al.*, 2024).

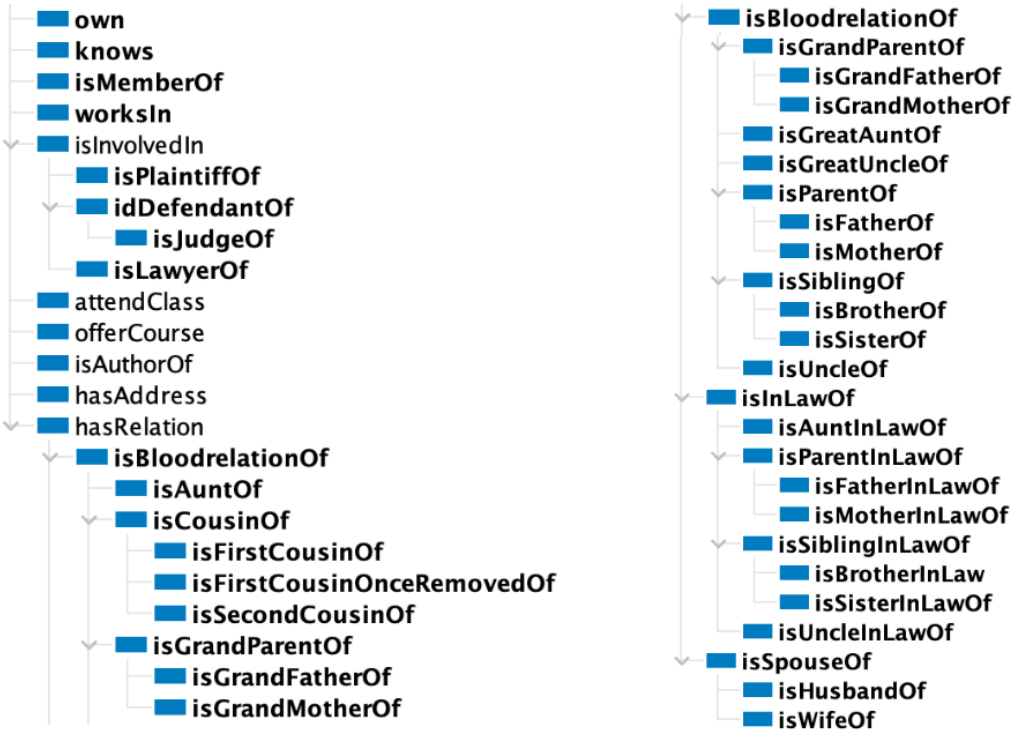

Figure 6.3: Object Properties (TORRES *et al.*, 2024)

6.4.1 Data Sample Mapping and Ingestion

To evaluate the proposed ontology, several Brazilian government databases were used, including RENACH, the Voter Registration Database, CPF, and RAIS, which report employment relationships between workers and employers (TORRES *et al.*, 2024). Data from Brazilian companies and various asset databases were also incorporated, including databases on vehicles, watercraft, aircraft, and real estate.

A sample of 1,000 individuals, drawn from the CPF database and matched with records from other databases, was retrieved (TORRES *et al.*, 2024). That data was thus entered into the ontology for assessment of its completeness with respect to classes, attributes, and relations. The evaluation confirmed that all selected data could be mapped to existing ontology elements, precluding the creation of new classes and relations and validating the ontology's completeness.

6.4.2 User-based Evaluation

The stakeholders in the ontology instantiation process identified several challenges (TORRES *et al.*, 2024). One of the major problems was merging data from similar entities across datasets, as most records lacked unique keys. For instance, data on legal proceedings in Brazil are published with the parties' names, making it challenging to precisely identify the entities due to homonyms. Another issue concerning homonyms arises when relational attributes are stored with the name of an entity alone. An example would be the "isMotherOf" relationship. When only the mother's name is recorded in someone's registration, differentiating between multiple individuals with the same name becomes difficult.

According to stakeholder feedback, these challenges arose from data inconsistencies rather than from faults in the ontology's knowledge model. Once proper data instantiations were made, the identified semantic issues were solved. During checking, there was no need for additional classes or relationships to emerge, confirming that the ontology fulfilled all requirements established during its specification stage, as in Table 6.11 (TORRES *et al.*, 2024).

Table 6.11: Application-based Evaluation (TORRES *et al.*, 2024)

User Requirements	**The way it was addressed in the ontology**
People who commit asset concealment fraud often buy high-value items such as automobiles, boats, aircraft, or real estate to convert liquid assets into physical assets that are harder to trace.	The proposed ontology maps more than a dozen types of assets fraudsters often use to hide their assets.
Fraudsters often place their assets in the name of relatives, such as parents, siblings, or spouses, to mask the true owner and avoid asset seizure.	Several relationship types, such as isMotherOf, isFatherOf, or isSpouseOf, can identify not only direct relationships, such as parents and children, but also indirect ones, such as grandparents, cousins, uncles, etc.

Continued on next page

Table 6.11 – continued from previous page

User Requirements	The way it was addressed in the ontology
People also tend to conceal assets by holding them with friends, neighbors, or close associates, further distancing the true owner from the property.	This type of semantics can be extracted from the proposed ontology by connecting people linked to the same address entity or who live close to each other, through address geocoding analysis. Furthermore, information such as people who are coursemates, wrote the same scientific article, participated in the same research group, or worked in the same company is an example of relationships in the ontology that indicate a possible close relationship among people.
Individuals hide assets by creating or using existing companies, often setting up complex ownership structures with distributed shareholding to obscure the true ownership. This fact can involve shell companies or multiple layers of corporate entities across different jurisdictions.	Mapping company data, especially partner data, helps meet this point (IS_PARTNER). In addition, the semantics associated with this type of relationship even allow the detection of a network of companies by traversing the graph.
Fraudsters often transfer assets to offshore bank accounts in countries with strict privacy laws, making it difficult for authorities to trace the funds.	This type of data is often restricted from government access or even omitted from statements made by people. Regardless, the proposed ontology is prepared to map this concept once the information is available (Asset), including allowing governments to follow the path of money.
Increasingly, people use cryptocurrencies to hide assets, as they can be more difficult to track due to their decentralized nature and the anonymity they offer.	As in the previous situation, these data are often restricted from government access or even omitted from statements made by people. Regardless, the ontology is prepared to map this concept (Asset/CryptoWallet) once the information is available, including allowing governments to follow the path of cryptocurrency.
Governments are among the primary victims of asset concealment fraud, despite commonly having databases containing information on individuals, their attributes, and their assets.	The proposed ontology enhances governments' capacity by proposing a centralized model for integrating dispersed data. It provides a holistic view of the problem, which is essential for analyzing and investigating this type of fraud.
Governments commonly lack a strategy for integrating distributed databases related to the same entities. In addition, similar data stored in different databases often exhibit inconsistencies.	This issue is addressed by constructing an ontology defining a model for integrating real-world concepts. This work also proposes methods to treat and correct possible inconsistencies in the dispersed data.

6.4.3 Application-based Evaluation

This application-based evaluation thus includes the ontological design within the environment of a Brazilian public agency, typically instantiated as a knowledge graph. The effectiveness assessment was based on its ability to assist the application's operational tasks. While theoretical evaluations focus on internal structure and logical consistency, this method aimed to measure different levels of applicability performance within a particular use case.

The ontology has been successfully put into production and integrated with the corporate systems at AGU. These enhanced systems were then made available to end users, who approved the outcome. The validation of the ontology attests to its absence of conflicting definitions or logical inconsistencies, enabling it to integrate smoothly into existing systems and data repositories. Automatic reasoning was also applied to detect inconsistencies, ensuring the ontology could perform its reasoning properly without encountering errors or contradictions.

Moreover, domain experts were involved in the assessment to confirm the ontology's ability to function across diverse systems without compromising data integrity and meaning in their exchanges. The successful integration and end-user approval indicated the ontology's usability and reliability in real-world scenarios, underscoring the original achievement in the application's design and deployment.

6.5 SUMMARY

This chapter presented the process of constructing the Asset Hiding Ontology (ACO) by applying the methodology proposed in Chapter 4. The ACO was developed by identifying clusters of investigative data and aligning them with existing semantic patterns. It was validated through a specific, data-driven, user-oriented process. In the end, the ACO evaluation demonstrated its effectiveness in formally representing the domain.

With the conclusion of this chapter, the semantic fragmentation gap was formally addressed. This ontology will serve as the unified schema for the next phase, Ecosystem Implementation, which will use the ACO to integrate and instantiate large-scale government data.

CHAPTER 7

ECOSYSTEM IMPLEMENTATION

After developing the ontology and knowledge graph construction processes, as well as the asset recovery domain ontology, it is time to validate the proposed ecosystem by creating a knowledge graph for asset hiding. To do this, we will execute the stages of the knowledge graph construction process presented in Figure 5.1, using the ontology already constructed and several government data sources, to which access was granted via a research project associated with a decentralized execution term between the University of Brasilia and the Attorney General's Office.

7.1 BUSINESS UNDERSTANDING

In this section, we will address the activities developed within the scope of the first macro process - Business Understanding, and its substeps: Defining Business Objectives and Requirements; Planning Data Governance; Identify and Obtain Access to Data; and Map entities, relationships, and Attributes.

7.1.1 Defining Business Objectives and Requirements

As previously discussed, the main business problem is the fragmentation of information across multiple, disintegrated government repositories. This gap limits the Attorney General's Office's ability to process, cross-reference, and analyze large-scale data. It also prevents the proactive identification of complex fraudulent schemes and favors the actions of debtors who squander their assets to thwart the recovery of public debts. Using a knowledge graph emerges as a viable solution to transform this reactive scenario into a proactive, data-driven model.

The business objectives guiding this project are:

1. **Centralize and Integrate Data**: Unify data from diverse government sources into a

single, connected, and semantic view, overcoming information fragmentation.

2. **Increase Proactivity**: Shift the investigative paradigm from reactive to proactive, enabling the AGU to identify asset concealment schemes in the early stages, before the assets are completely dissipated.

3. **Optimize Investigative Efficiency**: Reduce the time and manual effort required to map complex relationships between people, companies, and assets, automating the discovery of connections that currently require months of work.

4. **Generate Actionable Intelligence**: Go beyond simple data queries, enabling the discovery of hidden patterns and relationships, such as the use of "front men," shell companies, or complex corporate structures, which are difficult to detect with current tools.

5. **Create a Replicable Solution**: Develop a modular and agnostic computational ecosystem that, while focused on the AGU's use case, can be adapted to other public institutions with similar challenges.

The knowledge graph must be able to answer essential questions for AGU investigators, such as who the true beneficiaries of a given asset are, considering family, corporate, and trust ties; or whether there are networks of relationships (front men, shell companies, silent partners) that connect a debtor to undeclared assets.

The project's success will be measured through quantitative and qualitative performance indicators. Key success factors for this project include the percentage increase in funds recovered by the AGU in asset fraud cases, the reduction in the time required to complete a complex asset concealment investigation, the tool's adoption rate, and the number of integrated data sources.

7.1.2 Planning Data Governance

The construction of the Data and Knowledge Governance Summary Plan is a pillar of the project. It is designed to ensure reliability, security, and legal compliance throughout the graph's lifecycle. The plan covers four main areas: roles and responsibilities, quality policies,

legal compliance, and auditing. First of all, the following Roles and Responsibilities (RACI) were defined to ensure clarity and accountability:

1. **Knowledge Manager**: A team of AGU lawyers and analysts, experts in the asset recovery domain. They were responsible for curating and validating the graph's content and resolving data ambiguities. They are also responsible for ensuring that the inferences generated are legally relevant and accurate.

2. **Knowledge Engineer**: A technical team of UnB, responsible for designing, maintaining, and versioning the reference ontology and the KG. This role translates business requirements into a logical structure of classes, properties, and relationships representing the asset obfuscation domain.

3. **Data Steward**: This joint technical team (AGU/UnB) manages the technological infrastructure, data ingestion pipelines, platform security, and access control. It also ensured that the graph operates efficiently and securely.

4. **Governance Committee (Joint AGU/UnB)**: A strategic group composed of AGU managers and leading UnB researchers responsible for approving significant changes to the ontology, arbitrating highly complex data conflicts, and ensuring the project's continuous alignment with institutional objectives.

In a sequence, it was necessary to define the following quality and versioning policies: every entity or data ingested into the graph must be accompanied by essential metadata, including the source, the ingestion/update date, and the information's sensitivity classification; the ontology must follow a semantic versioning system, allowing schema evolutions to be tracked. About the legal compliance, the project must operate in strict compliance with Brazilian legislation, particularly the General Data Protection Law. Regarding security, data access must be strictly controlled, with read and write permissions defined according to users' roles and need-to-know. The Immutable Audit Trails mechanisms must be implemented to ensure transparency and security.

7.1.3 Identify and Obtain Access to Data

In total, we used 15 input databases from the initially proposed list in the process, including those related to the domain and whose access could be granted. The central databases used are listed in Table 7.1. The final analysis is presented in the following sections.

Table 7.1: Most Relevant Datasets used in this work

Category	Description
Biographical data of People	Individual Taxpayer Registry in Brazil - CPF Database, National Register of Licensed Drivers - RENACH, National Voter Registry - TSE Database, Death Control System - SISOBI
Taxes, Finances, Labor and Employment	Annual Social Information Report - RAIS, General Registry of Employed and Unemployed Persons - CAGED
Data from Companies	National Registry of Legal Entities - CNPJ
Address Data	Brazilian National Postal Code Dataset - CEP, National Address Registry for Statistical Purposes - CNEF
Data from Assets and Owners	The Brazilian Aeronautical Registry - RAB, National Motor Vehicle Registry - RENAVAM, Vessel Registry in Brazil, Registration and Salaries of Brazilian Federal Public Servants, Receipt of Funds by Beneficiary Database.
Others	Database of Environmental Violation Notices and Fines - IBAMA

7.1.3.1 Brazilian National Postal Code Dataset - CEP

In Brazil, postal codes consist of 8 digits, each representing a geographic division, as shown in Table 7.2. Thus, each Brazilian state is guaranteed one or more address number ranges, just as each city within a state is guaranteed a portion of the state's ranges, which is distributed among streets, or a single zip code of the town, depending on the city's population. Typically, only municipalities with more than 50,000 residents in urban areas have postal address codes by street (CORREIOS, 2025).

Table 7.2: Datasets Description.

Digit Position - X	Description
X0000-000	Region
0X000-000	Subregion
00X00-000	Sector
000X0-000	Subsector
0000X-000	Subsector divisor
00000-XXX	Distribution suffix

According to (CORREIOS, 2025), for municipalities with a single or generic ZIP code, the distribution suffix is usually "-000". For municipalities with street addresses, the distribution suffix ranges from 000 to 899, with 899 designating, with some exceptions, rural areas. The Brazilian Postal Code Company uses distribution suffixes 900-999 on special occasions. The range of 900-959 is commonly used for large mail recipients (condominiums, companies, institutions, etc.). The range from 960 to 969 is suitable for commercial promotions, and the range from 970 to 989 is specifically for Post Office Units that provide Post Office Box services. Large recipients in rural areas, each designated a Community Post Office Box, occupy between 990 and 998. Finally, for internal services of letters, cards, envelopes, and parcel replies, the distribution suffix 999 is used.

The postal code dataset consists of 1,504,736 records. Of these, 1,459,732 represent street-coded addresses, 20,250 were assigned to large mail recipients, 12,742 to physical post offices, and 2,165 to post office box locations. Additionally, the Brazilian government assigned 9,847 ZIP codes to locations without street addresses, of which 4,304 are districts, 4,473 are cities, and 1,070 are villages. Districts are an internal territorial and administrative subdivision of the Municipality, legally established, but do not have political autonomy; finally, villages are the smallest, being a simple population cluster of an informal or developing nature, without territorial limits or their own administrative status, depending directly on the headquarters of the District or the Municipality to which it is subordinate.

In addition to the ZIP codes used in each small municipality with only one ZIP code, the ZIP code database includes ZIP code ranges assigned to each state and municipality, coded by street address, and for many existing districts and villages. There are also specific zip codes for Physical Post Offices, Post Office Boxes, and Large Mail Recipients, which are typically

businesses, large residential communities, or government agencies that handle a high volume of mail and packages. This entire structure is distributed across distinct tables.

7.1.3.2 National Address Registry for Statistical Purposes - CNEF

The Brazilian Institute of Geography and Statistics (IBGE) designed the National Address Registry for Statistical Purposes (CNEF) as an official repository of addresses nationwide. Over time, the CNEF has transcended its primary role as logistical support for IBGE's data collections, establishing itself as a highly relevant and reference source of geospatial information for identifying and locating households nationwide.

One of the CNEF's most significant strengths for academic research and public policy planning lies in its granularity, allowing data aggregation across different territorial segments. Its ability to extract precise household numbers using the Postal Code (CEP) enables analytical potential for demographic, urban planning, and logistics studies, allowing the construction of population estimates and housing density analyses at a level of detail that few other data sources allow. The quantification of residences by Postal code, derived from the CNEF, offers relevant support for investigations that require understanding the spatial distribution of households in Brazil.

7.1.3.3 Individual Taxpayer Registry in Brazil - CPF Database

The central database for storing personal data in Brazil is the Individual Taxpayer Registry (CPF Database). It also stores the unique identifier of Brazilian citizens, which shares the same name as the database—CPF. Because of its importance, access to the CPF database is strictly controlled. Sensitive information associated with the CPF, such as personally identifiable data, is protected to prevent misuse and to ensure individuals' privacy. The data analyzed were for May 2024.

The CPF database includes a comprehensive set of attributes, grouped into related categories, that provide detailed personal, demographic, and administrative information about individuals. Personal Identification Data includes attributes essential for identifying individuals. These attributes are the CPF (the unique taxpayer registry number), the full name, the

birthdate, and the mother's full name, which are often used for identity verification. To further validate identity, the database includes the voter registration ID and, if applicable, the date of death.

The Address Data provides detailed location information for individuals and includes the street name, address number, and address complement, such as apartment details, the neighborhood, municipality, state, and postal code. For individuals' places of birth, attributes indicate the corresponding state and municipality in which the person was born.

The Contact Data includes attributes related to communication, such as the area code and the phone number, facilitating direct contact with individuals. Demographic and Foreign Residency Data captures additional details, especially for individuals outside Brazil. Further details include the name of the foreign country and the individual's nationality.

The Occupation and Employment Data group focuses on professional activity and work-related information. There is an attribute that describes the individual's primary occupation. The Administrative and Registration Data include critical information for managing CPF records, including when the CPF was initially registered and other attributes indicating the most recent update to the record and the CPF's current status (e.g., active or suspended).

7.1.3.4 National Register of Licensed Drivers - RENACH

The National Register of Licensed Drivers (RENACH) is the Brazilian government's primary database for licensed drivers. The agency responsible for managing and centralizing data is the National Traffic Secretariat (SENATRAN), which maintains a unified repository of data collected by each state's State Traffic Departments (DETRANs).

The use of a centralized architecture aims, among other things, to ensure the uniqueness of each driver's record and prevent fraud, such as the issuance of a new license in one state. At the same time, the individual has restrictions on another, and ensures the integrity and consistency of the driver's record nationwide.

Although the RENACH data collection contains information on a driver's entire life cycle, including training and violation data, the scope of this work included only registration and biographical information, such as full name, parentage, date of birth, address, Individual Taxpayer

Registry (CPF) number, and the driver's unique identification number (RENACH number).

7.1.3.5 National Voter Registry - TSE Database

The National Voter Registry, managed by the Superior Electoral Court (TSE), is an essential source of civil identification in Brazil. In addition to its primary function of identifying Brazilian voters, it is a centralized structure with widespread updates through the Regional Electoral Courts (TREs) in each state. In addition to the CPF identifier, the registry has a unique voter identifier linked to voter registration.

One advantage of this registry is that it typically has the most up-to-date address information. In Brazil, citizens participate in elections every two years, and voting locations are defined by their voter domicile. Another highlight was the implementation of biometric registration, which combined highly unique data with traditional registration, such as fingerprints of all fingers and facial photographs. This biometric layer increased the security and reliability of the registry, making it an effective tool in preventing identity fraud.

Despite grouping information from the candidate's electoral history, the data available for this work consisted only of basic identifying information, such as full name, date of birth, address, and affiliation.

7.1.3.6 Death Control System - SISOBI

The Death Control System (SISOBI) database is a registry of deceased individuals that centralizes death records throughout Brazil. The Ministry of Social Security manages the database, which is operated by the National Institute of Social Security (INSS). The database is linked to the need for Social Security oversight and auditing to mitigate fraud and improper payments of Social Security and welfare benefits to deceased beneficiaries.

Each record in the database is associated with an individual and contains a set of key variables for their unique identification. The main fields available in this database are personal identifiers, such as the CPF and Employee Identification Number, demographic data, such as the deceased's full name, gender, and date of birth, and death information, such as the date of

death and, sometimes, the date of death registration.

7.1.3.7 National Registry of Legal Entities - CNPJ

The National Registry of Legal Entities (CNPJ) gathers information on all legal entities operating in Brazil. The Brazilian Federal Revenue Service (RFB) owns this database, which has the primary function of uniquely registering and identifying each company, association, foundation, or other legal entity in Brazil, supporting tax administration, oversight, and economic regulation.

The database's data collection is granular and multidimensional, detailing the composition and characteristics of each entity. The database comprises three main entities: companies, establishments, and partners. For each company and its respective establishments (headquarters and branches), information such as the corporate name, trade name, address, opening date, registration status (active, inactive, etc.), and the National Classification of Economic Activities (CNAE) is provided. The partner database connects legal entities to individuals (via CPF) or other companies (via CNPJ), specifying each member's qualifications within the corporate structure and the date of incorporation.

The open data policy transformed this registry, previously primarily for internal government use, into a publicly accessible resource, except for the CPF identification number of individual partners. This fact means that anyone can directly import data and establish relationships between companies. However, additional processing is required because the unique identification number for individual partners is partially missing.

For complex investigations such as asset concealment, the ability to cross-reference partner and company data allows for mapping complex corporate structures, identifying the use of "front companies" and shell companies, and uncovering networks of relationships that would otherwise be impossible to observe. The interconnection between Individuals and Legal Entities is key to tracking the flow of resources and understanding how assets are distributed and hidden in corporate structures.

7.1.3.8 Annual Social Information Report - RAIS

The Annual Social Information Report (RAIS) is Brazil's central social and labor information storage database. It serves, among other purposes, as a basis for proving length of service for retirement, granting the Salary Bonus payment, and calculating the FGTS (Unemployment Fund for Severance Indemnity). It contains over 130 attributes that provide data on an individual's employment relationships.

Many of these attributes are essential for enriching user data, especially those that include identifiers, such as the employee's work card number and series, the employee's PIS (Social Security Income Tax) number, or even the profession they hold. Employment Relationship Data details the nature of the employment relationship. This category includes attributes such as the hire date, the termination date (if applicable), the Brazilian Occupational Classification (CBO) code describing the worker's role, the type of employment relationship, and the monthly remuneration.

Employer Address Data provides the location of the establishment where the worker works. This group includes the street, number, address, neighborhood, municipality, state, and postal code (CEP). This information is beneficial for monitoring and for regional statistics.

7.1.3.9 General Registry of Employed and Unemployed Persons - CAGED

Until 2002, the General Registry of Employed and Unemployed Persons (CAGED) was the Brazilian government's main instrument for monitoring and overseeing the formal labor market, recording hiring and dismissals under the Consolidation of Labor Laws. Its primary purpose was to provide monthly data for the preparation of labor statistics and to serve as a basis for granting the Unemployment Insurance program. Although its mandatory use was replaced by eSocial in 2020, its historical collection remains a fundamental data source.

Its attributes are valuable for enriching user data, mainly because they contain unique identifiers. For workers, these include the PIS/PASEP number and CPF. For employers, the CNPJ (National Registry of Legal Entities) or the Specific INSS Registry is recorded, allowing for a precise link between the individual and the organization.

Employment Relationship Movement Data details the dynamics of the employment rela-

tionship. This category includes attributes such as the date of hire, the date of termination (when applicable), the type of movement (first job, reemployment, dismissal with or without cause), the Brazilian Classification of Occupations (CBO) that describes the worker's role, and the hiring salary. Employer Location Data provides the location of the establishment where the worker works. This group of attributes includes the street, neighborhood, municipality, federal unit, and Postal Code.

7.1.3.10 The Brazilian Aeronautical Registry - RAB

The Brazilian Aeronautical Registry (RAB) is Brazil's central civil aircraft registry. The National Civil Aviation Agency manages this database. It serves as an instrument for identifying and controlling the national air fleet, for ownership registration, for encumbrances, for inspection, and for operational safety purposes. The database contains dozens of attributes that detail each aircraft's technical, legal, and operational characteristics.

Many of these attributes are essential for data enrichment, especially those that serve as unique identifiers. The most important is the registration mark (e.g., PP-ABC, PT-XYZ), which serves as the aircraft's "license plate." Other necessary identifiers include the aircraft's serial number, engine number, and propeller number, allowing precise component traceability.

This database also details the aircraft's fundamental characteristics and ownership, including manufacturer, model, year of manufacture, owner's name (individual or legal entity), and CPF or CNPJ. The registration distinguishes the owner from the operator, including the latter's identification data, which is vital for determining operational responsibility.

7.1.3.11 National Motor Vehicle Registry - RENAVAM

The National Motor Vehicle Registry (RENAVAM) is the database that integrates the records of all land vehicles in circulation in Brazil. Managed by the National Traffic Secretariat, its objective is to unify and track each car's history from its first registration to its final deregistration.

The RENAVAM code is the vehicle's primary identifier. It is a unique and unchangeable

number that accompanies the car throughout its existence. Other identifiers are the Vehicle Identification Number (VIN) and the license plate. There are also attributes related to the vehicle's technical specifications, such as make, model, color, year of manufacture, model year, make/type, fuel type, and horsepower. The owner is also an attribute that establishes the legal link between a person or company and the vehicle.

7.1.3.12 Vessel Registry in Brazil

The Brazilian Maritime Authority, under the Brazilian Navy, oversees vessel registration in Brazil. This database is the sector's main regulatory instrument. It grants nationality to vessels and ensures the legal security of their ownership and operation. The registry covers various types of vessels, each with a unique registration attesting to its compliance with navigational safety regulations. From this research's perspective, the Vessel Registry database provides essential information about an asset that is frequently used to conceal assets, including its approximate value and the owner's identification.

7.1.3.13 Receipt of Funds by Beneficiary Database

The open database on "Receipt of Funds by Beneficiary" allows citizens to track to whom, how much, and for what reason public funds from the Brazilian federal government are being paid. This database, maintained by the Office of the Comptroller General of the Union (CGU) and made available primarily through the Brazilian Transparency Portal, consolidates and details all direct payments made by the Federal Government to its beneficiaries. A "beneficiary" can be any individual, legal entity, or government entity that has received federal funds.

The essential fields that make up this database are, among other things, the period in which the payment was recorded, the hierarchical identification of the government entity that made the payment, the name of the individual or the corporate name of the company that received the funds, the unique identifier of the recipient (in the case of the CPF, this is partially anonymized), and the amount received. For this experiment, data were obtained from January 2023 to September 2025

7.1.3.14 Registration and Salaries of Brazilian Federal Public Servants

The CGU publishes a monthly set of government data on the Federal Government Transparency Portal, with the database of civil and military public servants of the Federal Executive Branch among its most important assets. This information repository provides a detailed overview of the federal government workforce. It allows for the consultation and download of microdata, including the employee's employment record and information on their employment relationship, position, assignment, and status (active, inactive, or retired). The database's contribution to this research lies in the granularity of the remuneration information provided. Monthly salary details, including base pay, severance pay, bonuses, and mandatory deductions, are available for each public servant.

This database is monthly and separated into 11 different files, one for each of the following categories: Central Bank retirees, Reserve Military personnel, other executive branch retirees, Central Bank pensioners, Ministry of Defense pensioners, and other pensioners, military personnel, Central Bank employees, and other public servants, legal fees paid to federal attorneys, and jeton fees paid to members of public company boards. Another issue is that the CPF data is partially missing, requiring a deduplication process to match the partial CPFs associated with the name to other entities in the database.

7.1.3.15 Database of Environmental Violation Notices and Fines - IBAMA

The Brazilian Institute of the Environment and Renewable Natural Resources (IBAMA) maintains the database of Environmental Violation Notices and Fines. This database contains information on administrative sanctions imposed for violations against flora, fauna, fisheries resources, and other environmental violations. Each record corresponds to a specific violation notice and stores the violation data, the geographic location, and the person responsible. This study used these data as input to build a model for discovering links between offenders associated with different violation notices.

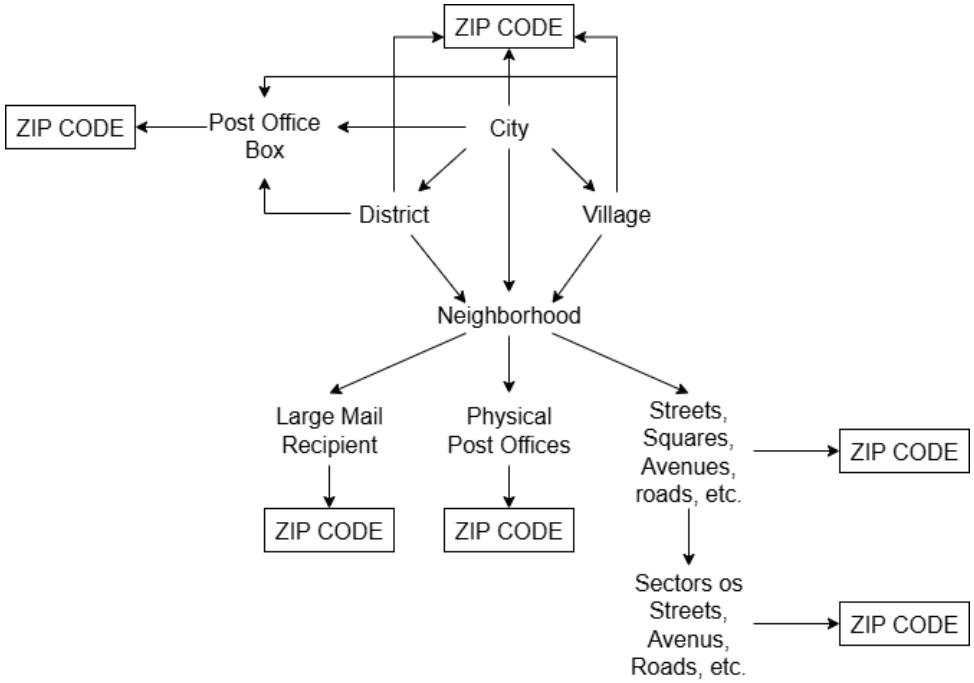

Figure 7.1: Brazilian Zip Code hHerarchy

7.1.4 Map entities, relationships, and Attributes

After the team has analyzed and defined the definition structure, the next step is to align the domain elements (entities, relationships, and attributes) with the project's base ontology. This stage is characterized precisely by mapping entities and relationships between the data and the domain ontology.

7.1.4.1 Brazilian National Postal Code Dataset and National Address Registry for Statistical Purposes

Although the reference ontology adequately supports addresses in the conventional structure used in most Brazilian government databases, it was unsuitable for incorporating the Brazilian postal system's postal database, particularly because of the semantics extracted from its attributes. After analyzing the dataset, it was possible to establish a new set of entities and their relationships, as shown in Figure 7.1.

Based on the attribute analysis, a change to the ontology was proposed and approved to incorporate the entities and relationships shown in Figure 7.2. It is essential to emphasize the relevance of this hierarchical structure, as it allows one to understand the proximity between

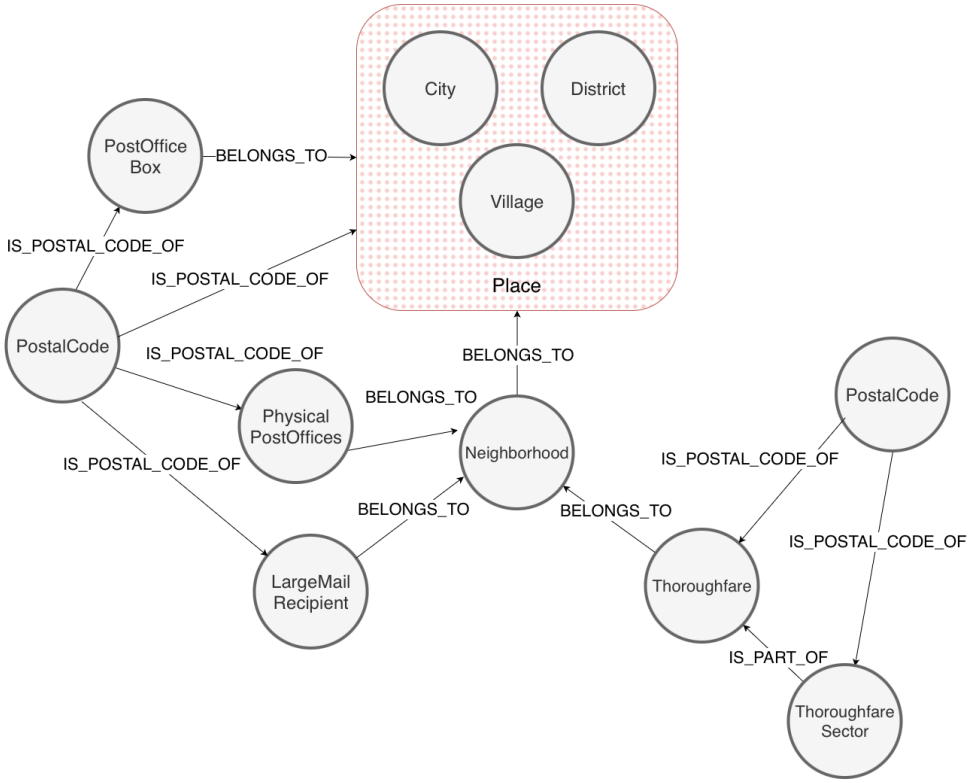

Figure 7.2: Brazilian Zip Code Structure

two Postal codes simply by instantiating them in the ontology and navigating their structure, without the need for georeferencing. Two people with different Postal codes may be on the same street, in other sectors, in the same neighborhood, in the same district, or in the same city. Each view of this structure brings its own concept of proximity.

In the context of this work, the CNEF database was included not to provide entities or relationships, but to aid in the construction of the heuristic for calculating social proximity between two people, which will be detailed in a later section. To this end, entities modeled as postal codes, obtained from the analysis of the Brazilian postal address code database, were enriched with the number of residences.

7.1.4.2 Individual Taxpayer Registry in Brazil - CPF Database

Some database attributes can be directly mapped to entities in the ontology, such as Person, Email, Address, Postal Code, and Phone Number, as shown in Figure 7.3. It is also possible to perform a direct correlation of all relationships in Figure 7.3, except IS_MOTHER, since, despite having an attribute with the person's mother's name, the existence of homonyms in the

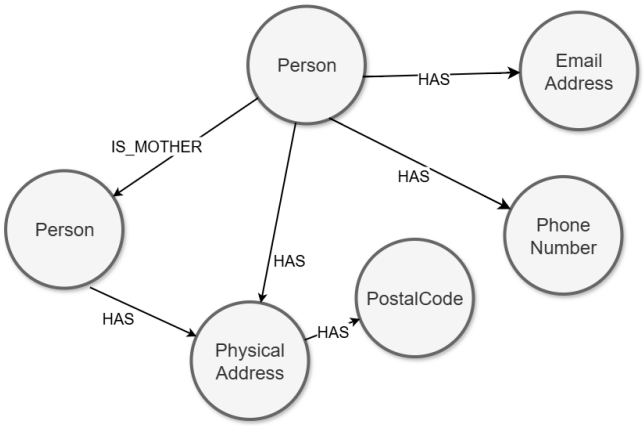

Figure 7.3: CPF Database Structure

database prevents a direct correlation. A link discovery model will be proposed for this specific case in a subsequent section.

Another critical issue is that Postal code correlation directly affects address linkage. Using the secondary association of links created through the Postal code database, it is possible to infer proximity between people with different Postal codes, even without geocoordinates. Finally, it is essential to emphasize that, in future work, a record linkage model for address records will be developed to address problems arising from errors in users' entering address data.

7.1.4.3 National Register of Licensed Drivers, National Voter Registry and Death Control System

Since the project uses data from the CPF database as the master data for individuals, RENACH's role is primarily to collect potential new residential addresses and establish family ties, under the same conditions as the CPF database, since the occurrence of homonyms prevents direct generation, as shown in figure 7.4

Considering again that the project uses the Brazilian Individual Taxpayer Registry Identifier (CPF) as the primary key for the unequivocal identification of individuals, the Superior Electoral Court (TSE) database is used as a complement. Its function is to aggregate relevant information, such as additional residential addresses, and, most importantly, to infer possible filiation links. The prevalence of homonyms, however, makes the automated creation of these relationships unfeasible, requiring analysis for correct linking, as detailed in the figure 7.4.

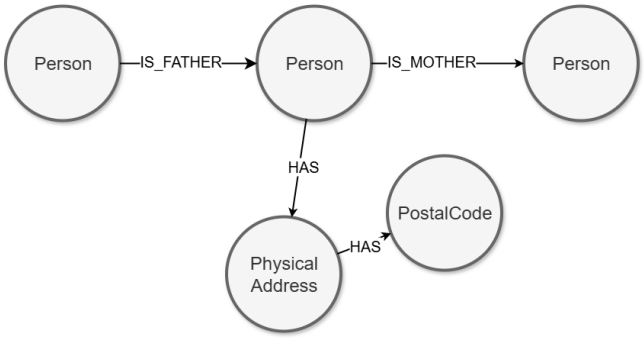

Figure 7.4: RENACH and TSE Database Structure

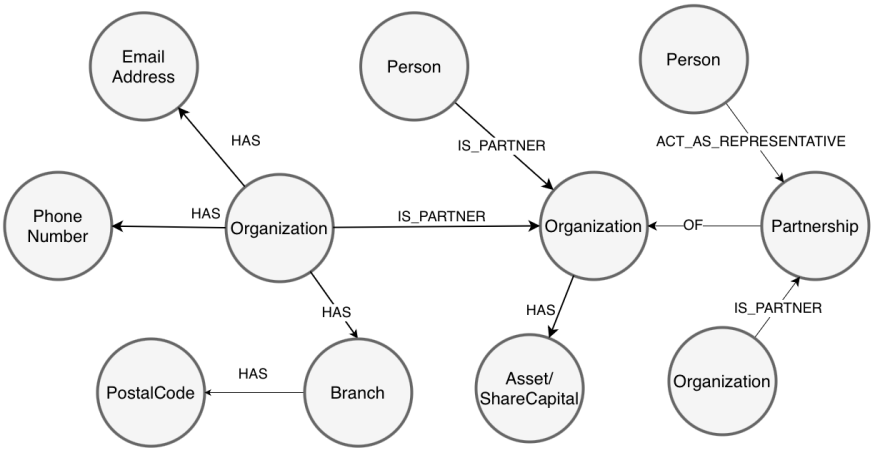

Figure 7.5: Companies Structure

The SISOBI database was used to enrich personal records with death information. This attribute proved essential for the training sets that enabled the heuristic for identifying kinship.

7.1.4.4 National Registry of Legal Entities - CNPJ

This dataset contains important attributes for identifying hidden links between individuals through shell companies. In Brazil, a company can have many branches, each stored in a separate table. Using the reference ontology, each branch was modeled as a Branch entity, and a company's share capital was modeled as an Asset.

The database allows direct linking of companies that are partners of other companies, but not for individuals who are partners of companies, as it does not provide a unique identifier for this type of partner, only a partial CPF number. To enable this linking, a heuristic was needed to deduplicate individuals based on their names and partial CPF numbers. The complete modeling structure used for the attributes of this database is shown in Figure 7.5.

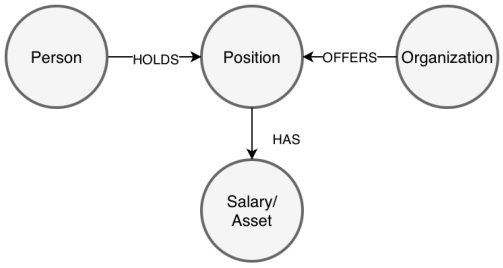

Figure 7.6: RAIS and CAGED Structure

Furthermore, during database analysis, we realized that when a company is a partner of another, a person may be named as its legal representative. To address these cases, the technical team requested a new adjustment to the ontology, which the governance committee approved. After that, an additional entity, Partnership, with two relationships, was added to link the partner company and its representative, using a unique structure.

7.1.4.5 Annual Social Information Report and General Registry of Employed and Unemployed Persons

The RAIS database modeling led to a change in the reference ontology's domain. Initially, the entity Person was linked to an entity Company via an employment relationship. The Person was also linked to another entity, Salary, via an ownership relationship. Therefore, it was impossible to determine whether there was any relationship between a given Salary and a Company when the Person had more than one employment relationship, even if inactive.

For this reason, a new change to the ontology was requested to add a new entity, defined as Position, and to determine a three-way relationship with it: with the Person who holds this job, with the employing Company, and with the Salary paid to the user based on this job, as detailed in Figure 7.6. CAGED data fits into the same context as RAIS, with similar attributes and sharing the same entity and relationship structure.

7.1.4.6 The Brazilian Aeronautical Registry, National Motor Vehicle Registry, and Vessel Registry in Brazil

The modeling of these three databases followed the same structure, with the assets - aircraft, watercraft, and motorcraft, which may have ownership associated with people or organizations,

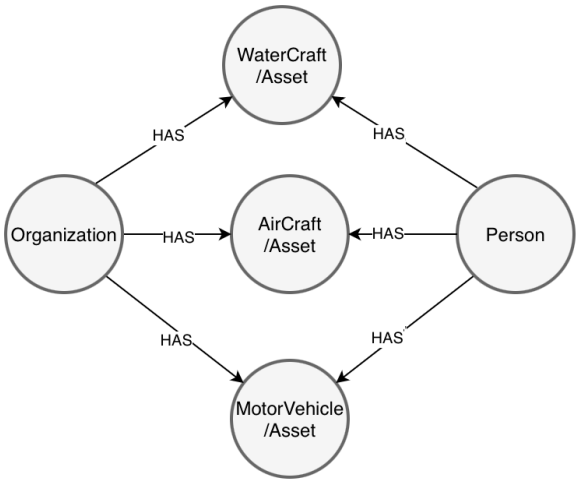

Figure 7.7: AirCraft, MotorVehicle, and WaterCraft Structure

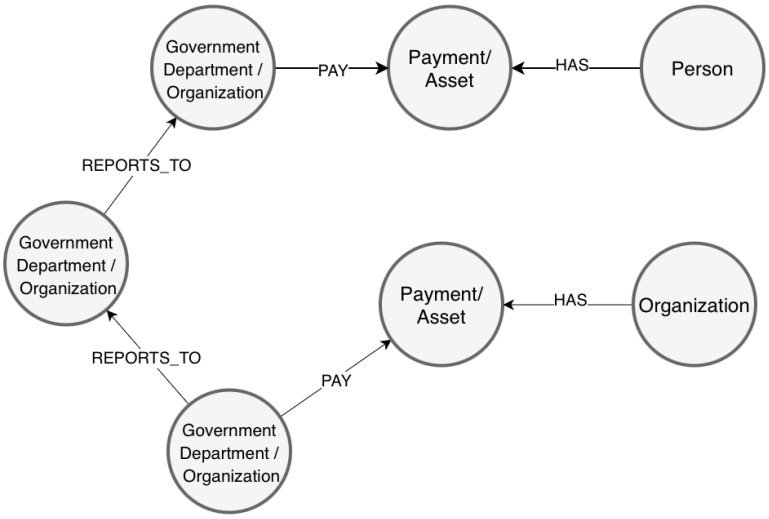

Figure 7.8: Receipt of Funds by Beneficiary Structure

as shown in Figure 7.7.

7.1.4.7 Receipt of Funds by Beneficiary

Like other open data in Brazil, records of individuals receiving funds are partially anonymized and must undergo deduplication. Links to organizations are made directly, since companies' unique identifiers are public. Another adjustment to the reference ontology was necessary, with the addition of a new entity, Government Department, representing the federal government department responsible for disbursing funds. The Brazilian government structure is hierarchical, and this hierarchy of government departments is also reflected in the final ontology (7.8).

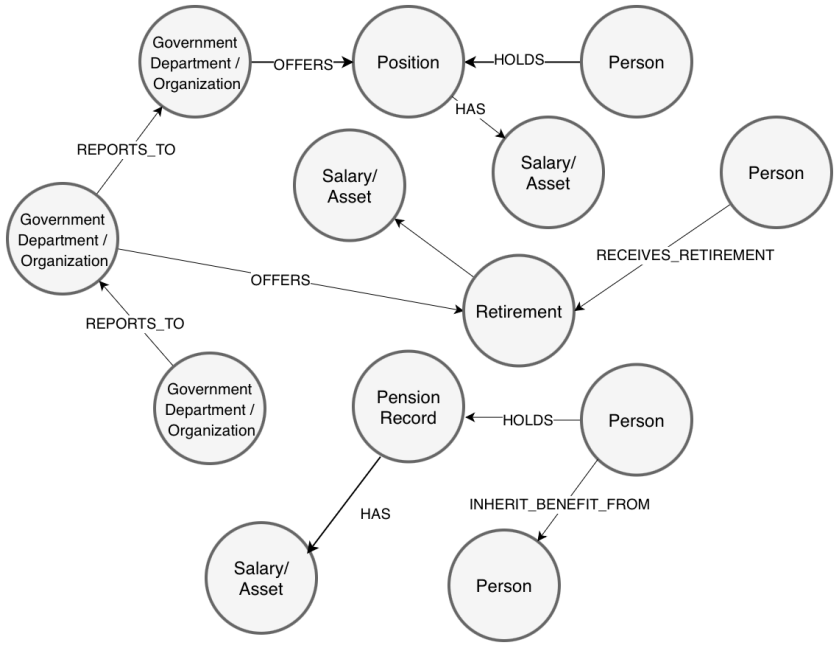

Figure 7.9: Salaries of Brazilian Federal Public Servants Structure

7.1.4.8 Registration and Salaries of Brazilian Federal Public Servants

The database containing the civil servant registry in Brazil is one of the richest available as open data. Besides providing important information about one of the main assets associated with an individual—their salary—it also provides a series of information about their ties to government departments and, in the case of pension payments, the pension recipient's family ties to the pension provider, as shown in Figure 7.9).

7.2 ENVIRONMENT PLANNING AND DEPLOYMENT

After the Business Understanding phase, the technological infrastructure that will serve as the basis for the knowledge graph must be planned and deployed. This step aims to ensure that the selected architecture, tools, and computing resources align with the needs identified by domain experts and support the subsequent process activities.

7.2.1 Define the technological architecture

The next stage was to plan a multi-tiered technology architecture to efficiently manage the entire knowledge graph lifecycle, from raw data ingestion to graph-based query execution.

The ecosystem is orchestrated primarily by scripts developed in Python and the AirFlow tool (ALBUQUERQUE *et al.*, 2025). Python scripts comprise all the developed data pipelines, from processing raw data for loading into the dimensional database to performing analyses and queries on the dimensional database itself and then preparing entities and relationships for loading into the graph-oriented database, the Knowledge Graph's storage location. AirFlow controls the execution of these scripts. StarRocks (LF PROJECTS, 2025) was used as the dimensional storage tool, and the Neo4J (NEO4J Inc., 2025) database was used for KG storage.

Python was selected for script creation due to its versatility and extensive library ecosystem, which makes it suitable for coordinating data flow across different systems and performing specialized tasks. Python was also used to build entity-deduplication, matching, and record-linkage models. Apache Airflow was selected as the main orchestration component due to its maturity, platform neutrality, and number of integration components.

StarRocks, an Online Analytical Processing (OLAP) database, was chosen as the intermediate data repository due to its ability to execute complex analytical queries (e.g., aggregations, joins, filters) on large volumes of tabular data. In the proposed architecture, StarRocks was primarily used as a centralized "staging area," where data was cleaned, standardized, and enriched before conversion to the graph structure.

Finally, Neo4J was chosen as the Native Graph Database because it offers a free access version, is an industry leader, and has a simple query language that allows even non-specialized users to perform graph searches. The Neo4j platform also includes a library of graph algorithms (e.g., PageRank, community detection, and shortest path). A high-level view of the proposed architecture is presented in Figure 7.10.

7.2.2 Deploy the Technological Environment

As Albuquerque described, the orchestration and automation layer of the ecosystem was implemented by installing two dedicated instances of Apache Airflow on separate servers (ALBUQUERQUE *et al.*, 2025) and two Neo4j instances. This environment was provisioned on four virtualized servers with the following specifications: one Xeon Silver 4114 CPU @ 2.20GHz, four vCPUs, and 16GB of RAM, two for development and homologation, and another two for

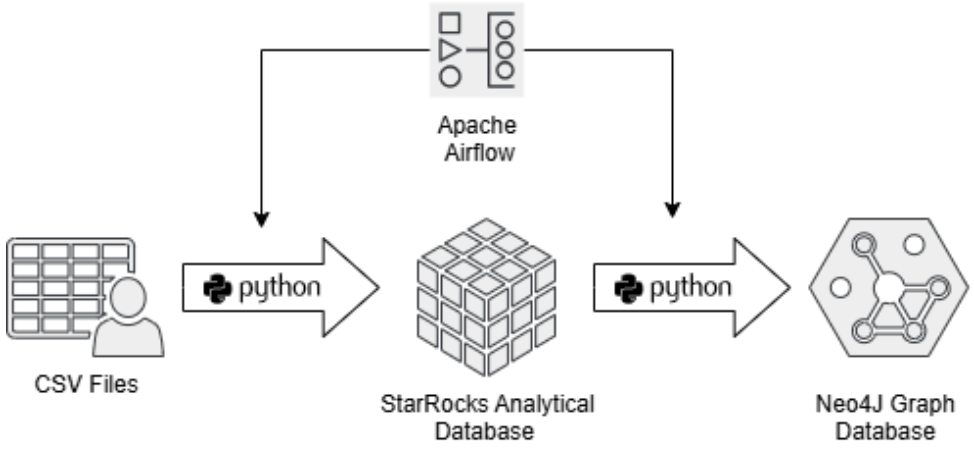

Figure 7.10: Technological Architecture

production. The primary function of Airflow instances was to automate and manage the entire data lifecycle that feeds the Knowledge Graph through directed acyclic graphs (DAGs). The routines, coded in Python, handled the whole data pipeline, from the ingestion and preparation of raw data to the extraction of entities and relationships, and the final loading into the Knowledge Graph in Neo4j.

Due to contractual restrictions, the Attorney General's Office chose the Neo4j Community Edition as its database. To mitigate the risks associated with the lack of a native high-availability cluster, the deployment strategy relied on data replication across two independent, isolated server instances. The first instance was defined as the main production environment. It was responsible for receiving all data loads and serving application queries. The second instance operated as a contingency and validation environment. It was maintained as a replica for backup and to validate new queries and data models before they were promoted to the main environment.

7.3 KNOWLEDGE ACQUISITION

The phase began with creating pipelines for data ingestion and preparation. The data sources underwent a four-step process: collection, preprocessing and transformation, verification and documentation, and loading into the staging area. From this structured repository, Python routines were used to extract entities, relationships, and attributes directly—without using Machine Learning—and load them into the Neo4j graph database, according to the ontology's

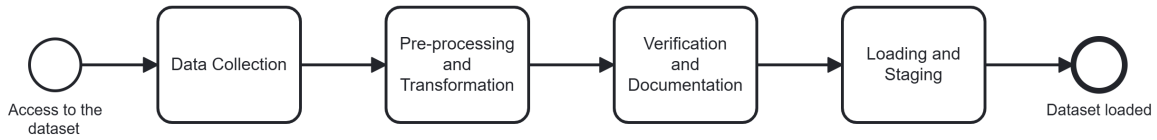

Figure 7.11: Ingest and Prepare Process

mapping.

7.3.1 Ingest and Prepare Data

This process has four main activities applied across all data collection, preparation, and ingestion pipelines (Figure 7.11). The initial phase of the process involves collecting data from previously mapped sources with previously granted access. Although sources can vary significantly in structure, the sources obtained in this study were in CSV, JSON, or positional format. Restricted-access sources were provided under confidentiality agreements, while open data sources were obtained directly from government agency portals.

Next, the data preprocessing and transformation stage begins, during which the collected raw data is subjected to a set of data engineering procedures to ensure its quality and adequacy. This step includes data-cleaning tasks to address duplicate records, inconsistent or null values, and noise. At the same time, data transformation adapts the source data's schemas and formats to align with the requirements of the target database's ingestion tools and pipelines. Data governance policies guide the execution of all these activities.

The third phase, data verification and documentation, serves as a quality control point, where the processed data is audited to ensure adherence to minimum requirements for completeness, consistency, and compliance with business rules. Significant anomalies or standard deviations identified during this verification can trigger a feedback loop, prompting adjustments to the graph ontology or a review of the governance plan.

At the end of this process, the Data Loading and Staging stage formalizes the transfer of data from the processing environment to the target repository. The data, now preprocessed, validated, and adequately documented, is loaded directly into the OLAP database, which will serve as the basis for analysis to extract entities and relationships.

This process also involved a basic record deduplication step. Since open databases only

contain partial information about the CPF identifier, it was necessary to cross-reference this attribute, along with the individual's name, with the individual database to obtain the full CPF number. Initially, we applied the same CPF mask used in open data to the CPF attribute in the tb_person source database to enable the comparison. Furthermore, only those records whose full name and partial mask were unique in the tb_person source database were retained in the cross-reference. Therefore, 269,686,751 of 269,821,454 met this criterion (99.95%).

The next step was to validate this strategy. To this end, data on active public servants was cross-referenced with data from this reference database for deduplication, resulting in the deduplication of 947,021 of the 960,855 existing records (98.56%). We observed that 13,834 records do not match between the databases. It was due to minor spelling errors, inconsistent use of abbreviations across databases, or even partial name changes due to marriage or divorce. To address this issue, we also tested comparisons using partial names. We began the tests using only the first five letters of the name associated with partial CPF to compare the two databases.

The problem with this strategy is that the number of unique elements in the reference database drops significantly—only 150,231,812 of the total 269,821,454 people can be deduplicated using partial CPF and the first five letters of the name. After a series of tests with varying character counts, the best result was the one using the first 12 characters of the name, which yielded a possible deduplication of 260,947,670 of the 269,821,454 records, and managed to reduce the number of undeduplicated records from 13,834 to just 5,801 after a second deduplication run. It increased the deduplication percentage of the original data set for active public servants to 99.50%. After validation, this strategy was applied to all open datasets that contained only partial information about the person's CPF.

7.3.2 Extract Entities, Relationships, and Attributes

Based on the imported data and the adjusted ontology, entities and relationships were incorporated into the knowledge graph in the Neo4J tool using Python routines. A view of some entities, with only identifiers for privacy reasons, and relationships is presented in Figure 7.12. The process imported entities across 30 classes. Some entities are associated with

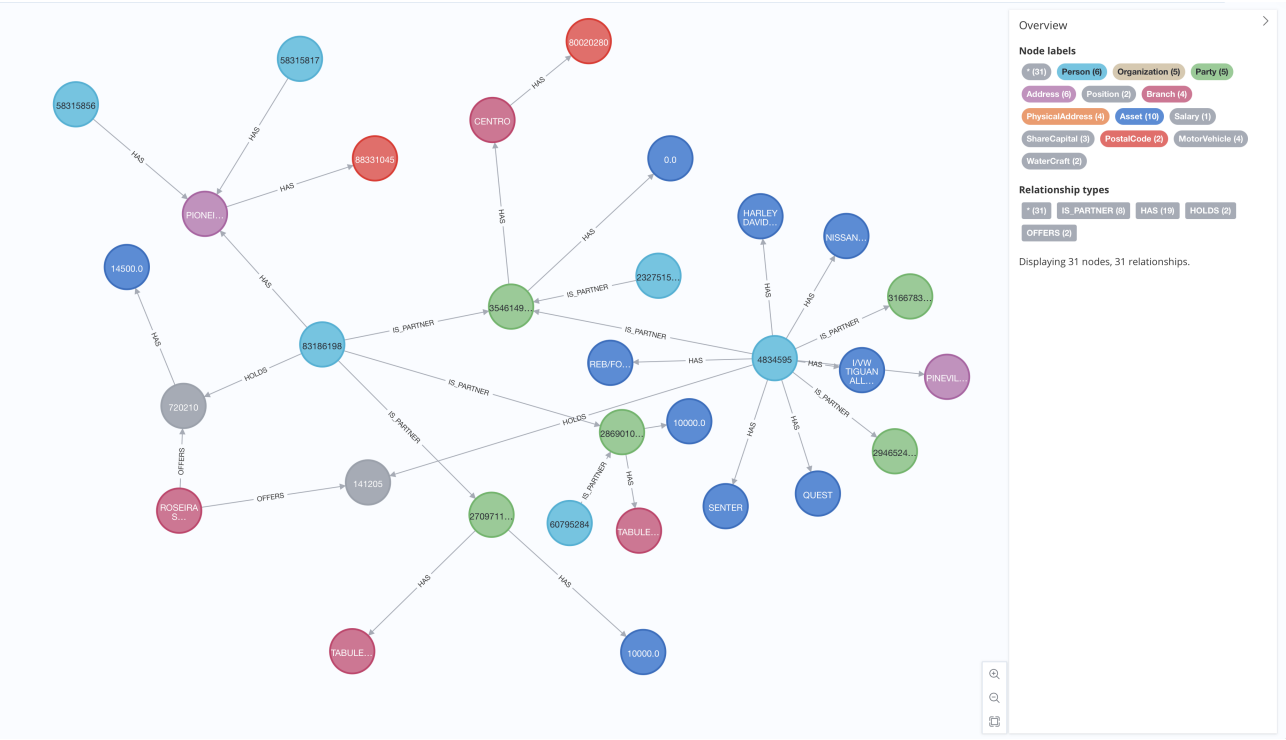

Figure 7.12: CNPJ Entities and Relationships

multiple classes in a multilabel structure. Among these entities, some relationships were also defined, such as: ACT_AS_REPRESENTATIVE, BELONGS_TO, HAS, HOLDS, INHERITS_BENEFIT_FROM, IS_PARTNER, IS_PART_OF, OF, OFFERS, or PAYS.

7.4 KNOWLEDGE IMPROVEMENT

7.4.1 Discover Entities and Relationships

This phase focuses on automatically identifying, using available data, instances, classes, properties, and links that were not mapped in the previous phase. Machine learning and natural language processing algorithms were applied to extract information that was not evident but vital to the investigative context.

7.4.1.1 Discover Kinship

In cases of asset concealment, family members are often used to mask the actual ownership of real estate, bank accounts, and other financial instruments (SHARMAN, 2011). In most cases,

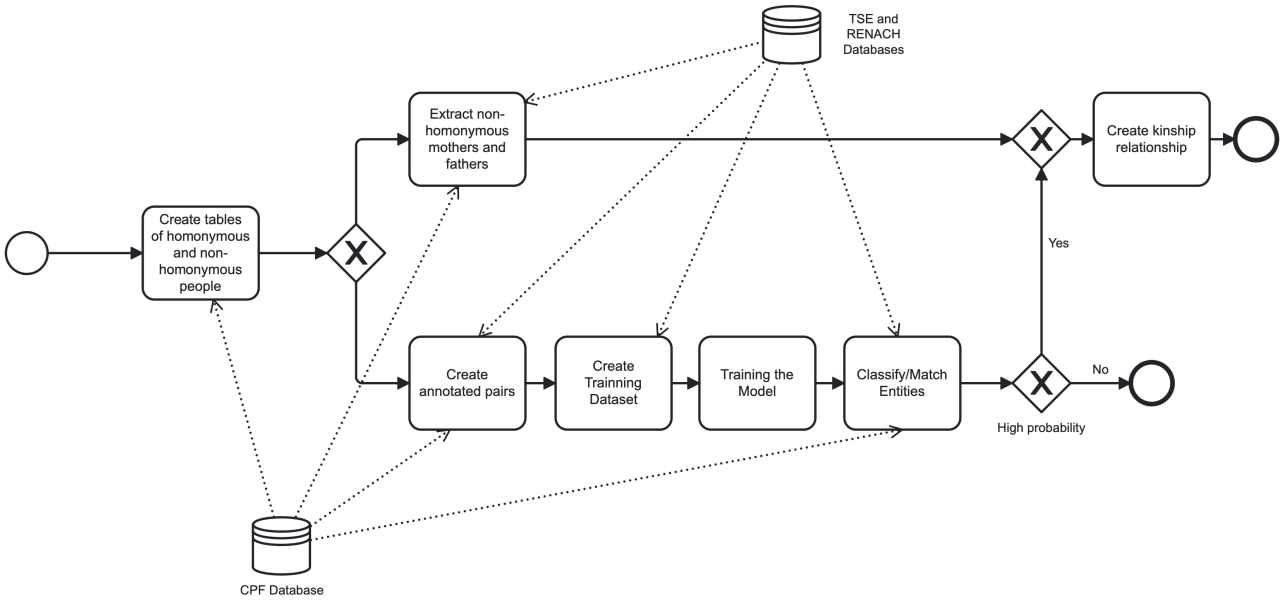

Figure 7.13: Process to deduplicate mothers and fathers

it is impossible to directly identify family ties from biographical data because a person's records do not store a unique identifier for their father or mother. Only one parent's name is often stored to identify an individual's record and avoid homonyms (MINISTRY OF JUSTICE OF BRAZIL, 2015). The presence of homonyms in person databases also prevents the automatic linking of parents' names to the personal registration database. To address this problem, this study proposed a machine learning model to identify kinship, developing a strategy based on correlated attributes to select fathers or mothers in cases of homonymy correctly. The steps necessary to execute this task are presented in Figure 7.13.

a) Create tables of homonymous and non-homonymous people

The first step involved partitioning the citizen database into two sets: individuals with unique names (non-homonymous) and those with names that appear at least twice (homonymous). The records identified in each category were stored in specific tables. The CPF database was used as input data for this activity, as it serves as a sufficient identifier for citizens in Brazil (BRAZIL, 2023). As a result of this step, 148,774,542 names were defined as unique in the CPF database. On the other hand, 10,448,580 person names appear twice in the database. These two tables - homonymous and non-homonymous- are essential in the subsequent steps since the unique parent names allow for the deterministic linkage to their unique identifiers when their names are unique in the database. On the other hand, the homonymous dataset enables the generation of possible parent-child pairs, which are used as input to the classification model

that identifies true relationships and discards incorrect ones.

b) Extract non-homonymous mothers and fathers

The next step was to match the mother's and father's names with the non-homonym table. This operation identified 92,643,243 possible mothers and 39,997,637 possible fathers. At first, matching only parents' names to a table where the parents' names are guaranteed to be unique would be sufficient to ensure an exact association between parents and children, using only the names without a unique identifier. However, no current database contains records of all the citizens of a country, especially since computer technology is relatively recent and may therefore lack records of people born before its implementation. Consequently, it is probable that some individuals' parents, especially those born closer to the database's launch date, are not registered. If these unregistered parents have the same names as some non-duplicate records, their children would automatically be associated with incorrect kinship records.

To mitigate this problem, an additional activity was included in this stage to avoid, at least, the creation of impossible relationships. Through the attributes of birth and death dates, it was possible to validate the child-parent tuples, preventing the association of records of parents who were born after the possible children, or of parents who died before the birth of the potential child, or even parents who were alive at the time of the potential child's birth but were not of reproductive age at that time. In this approach, mothers and fathers must be at least 10 years old, and mothers no older than 50 at the time of the probable child's birth. Additionally, they cannot have a date of death, in the case of those already deceased, before the probable child's birth date. After this validation, 89,912,294 mothers and 38,724,389 fathers were linked with their children.

c) Create annotated pairs

While the process in step (b) successfully identified a subset of high-confidence links, it cannot resolve cases involving homonymous parents. To address this, the next stage focuses on creating a labeled dataset to train a supervised classifier for this more challenging scenario.

To create the annotated pairs between a child and possible mothers and fathers, we matched the mother records in the CPF table and the father records in the TSE table against the homonym table generated in the previous step. This action generated 20,946,046 pairs between

people and possible fathers, and 29,128,600 pairs between people and possible mothers. In this context, because the homonym table contains only two records, for each associated person, two pairs were generated: either in the MATCH with fathers or in the MATCH with mothers.

Next, determine which record in the pair is true and which is false, given that a person can have only one father or one mother. To this end, we propose a rule to validate/exclude the possibility of a mother or father based on the children's birth and death dates and the possible fathers and mothers, similar to the rule used to validate relationships in the previous step. The difference is that they will be marked as non-fathers or non-mothers rather than having records that violate the validation rule disregarded. If both records in the pair meet the validation rules, both will be discarded and not included in the final training set. It is essential to highlight that, before this rule was defined, a group of people tried to annotate the record pairs manually. However, they had difficulty identifying elements to make the annotation.

Table 7.3 provides an example of the automatic label annotation process for a given pair. Note that for a given person, two possible parents were defined. However, based on the birth dates of both, it is possible to determine with some certainty which pair should be defined as parent and which as non-parent, since one of the potential parents would have been only 5 years old at the time of the supposed child's birth, which is scientifically impossible.

Table 7.3: Examples of Pairs Child-Father.

Father Name	**Child CPF**	**Father CPF**	**Father BDate**	**Child BDate**	**Label**
CA*******MOS	49********0	12********9	1961-04-XX	1941-09-XX	TRUE
CA*******MOS	49********0	23********2	1961-04-XX	1956-02-XX	FALSE

After executing this activity, 4,768,898 pairs of fathers and possible children remained, half labeled true and the other false, and 8,383,102 pairs of mothers and potential children, half labeled true and the other false.

d) Create Training Dataset

To construct the training dataset for the kinship discovery model, we employed a detailed cross-referencing approach linking children's registration attributes to their potential fathers and mothers. The central objective was to transform the relation between the characteristics of each pair (parent-child) into a similarity index. The similarity of the strings (texts) was

quantified using the Jaro-Winkler Distance metric, an adequate measure of how similar two character sequences are, even in the presence of minor typos or variations.

For the state of residence, state of birth, and postal code fields, the similarity score was replaced by a binary classification, assigning a value of 1 (one) for identical texts and 0 (zero) otherwise. It was done because the state fields consist of only two letters, and changing a single character maintains a high similarity index, even if the states differ entirely. With postal codes, due to their associated semantics, changing just one element can often lead to significantly different locations, even while maintaining a high similarity index.

The attributes Type of Public Place (Street/Avenue), Public Place Name (Street Name), House/Building Number, Neighborhood, City of Residence, State (UF) of Residence, Postal Code (CEP), Birthplace State and City (UF/Municipality), Birthplace City Name, combined with the predefined LABEL from the previous step, were used in creating the feature boolean matrix between parents and children, such as presented in Table 7.4.

Table 7.4: Sample Dataset for Kinship Classification

Mother CPF	**Child CPF**	**LABEL**	**Similarity Features**								
			Sim. Feat. Birthplace UF	Sim. Feat. Birthplace City	Sim. Feat. Street Type	Sim. Feat. Street Name	Sim. Feat. House Number	Sim. Feat. Neighborhood	Sim. Feat. City	Sim. Feat. State (UF)	Sim. Feat. Postal Code (CEP)
60********0	19********9	True	1.0	1.000000	1.000000	0.529762	0.000000	0.602564	0.411111	1.0	0.0
60********0	48********4	True	0.0	0.000000	1.000000	0.601190	0.000000	0.430556	0.447619	1.0	0.0
60********0	05********0	False	0.0	0.000000	1.000000	0.597222	0.000000	0.416667	0.658333	1.0	0.0
60********0	31********9	True	1.0	1.000000	1.000000	0.481481	0.611111	0.345238	1.000000	1.0	1.0
60********0	21********7	True	1.0	1.000000	0.000000	0.505556	0.000000	0.430556	0.447619	1.0	0.0

e) Training the Model

The previous step converted the relationship verification problem into a classic binary classification problem by transforming the record comparison into a vector of similarity features. The feature vector based on attributes presented in Table 7.4 is composed of similarity variables, and the target variable y has two possible values: TRUE or FALSE. Given this data structure and the need for performance and interpretability, the model evaluation was planned to include a set of classification algorithms.

To establish a baseline and ensure process transparency, Logistic Regression (HOSMER *et al.*, 2000) and Decision Tree (QUINLAN, 1996) were included. We chose Logistic because it provided coefficients that can demonstrate the statistical contribution of each binary attribute to the probability of the relationship. We selected the Decision Tree algorithm to provide interpretable results because it allows direct visualization of the logical classification rules.

Additionally, we used Bernoulli Naive Bayes (MURPHY *et al.*, 2006) for its inherent suitability and computational efficiency in handling binary features.

Regarding prediction performance, the analysis focused on ensemble models, namely Random Forest (BREIMAN, 2001) and Gradient Boosting (XGBoost) (CHEN; GUESTRIN, 2016). Random Forest is highly resistant to overfitting and efficiently handles high-dimensional binary features. XGBoost, in turn, is recognized for frequently achieving the highest accuracy on structured data, making it essential for accurately identifying the hierarchical importance of the most relevant features for determining kinship.

An essential feature of the proposed approach is its assessment of the probability threshold. In the fraud investigation problem at hand, it is more important to correctly recover the link (precision) than to collect all possible links (recall). Therefore, while the standard classification method uses a probability threshold of 0.5 to prioritize overall accuracy, practical application requires a high degree of certainty in classifying as TRUE (valid kinship) to avoid critical errors in data consolidation.

For this reason, the accuracy of the models that provided TRUE classifications was evaluated at three probability thresholds: 0.9, 0.95, and 0.99. This multifaceted analysis of the thresholds enabled us to determine the optimal model that balances recall and precision for each confidence level required by the final application, with precision prioritized.

The training sample used 5,867,581 samples, and the test sample used 2,514,67. The best result from the evaluation of each of the aforementioned algorithms, associated with the different probability levels, is shown in Figure 7.14. The result with the best accuracy, 99

Even with the focus on precision, the recall value was extremely low in this case. For this reason, even with the emphasis on precision, we chose a more balanced model: a Random Forest with a 0.9 probability threshold, achieving 97

f) Classify/Match Entities

Initially, pairs of possible mothers and children and possible fathers and children were assembled. Parents with 2 to 10 homonyms were used in the matching process, generating 181,825,473 possible mother/child pairs and 122,396,260 possible father/child pairs. Both sets were classified using the Random Forest model described in the previous step, yielding 18,915,785 preliminary

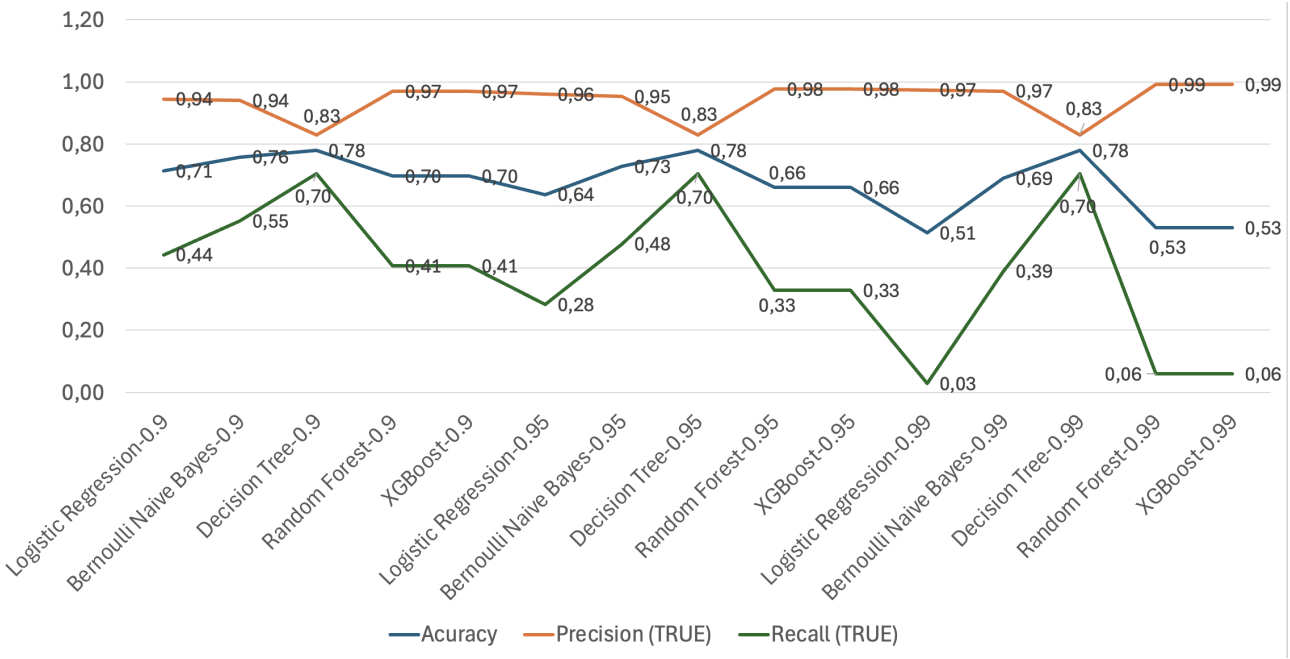

Figure 7.14: Kinship

mother-child links and 11,595,980 preliminary father-child links.

The final phase of the classification stage attempts to reduce false positives by removing records of children linked to more than one father or mother. In this stage, 1,916,589 provisional father-child links and 1,844,484 provisional mother-child links that violated the rule were removed. The final result generated 17,071,301 definitive mother-child links and 9,679,391 definitive father-child links.

The training phase lasted about 7 minutes, and the classification of all pairs took about 327 minutes, on a server with 64 GB of RAM and 8 processors, an acceptable time for the number of pairs analyzed (more than 300 million).

This experiment scientifically validates the capacity of supervised machine learning models to infer biological relationships solely from non-unique administrative attributes, effectively overcoming the absence of direct identifiers in legacy databases. Methodologically, it demonstrates that a Random Forest classifier trained on a dataset of homonymous pairs provides a robust, scalable mechanism for entity resolution, achieving high precision in a domain that typically requires manual verification.

7.4.1.2 Social Proximity

One topic involving address geolocation in public security is the fight against asset concealment fraud. In this type of crime, one way to conceal assets is by transferring ownership to a neighbor or close associate. Investigators often have to cross-reference large numbers of people's records and their addresses scattered across multiple databases to identify this type of correlation (TORRES *et al.*, 2024). The problem is that manually handling this scattered data is not feasible.

One possible solution would be to use automation to calculate the physical distance between two addresses via geocoding and, based on this information, identify a potential relationship between the users. The primary challenge of applying geocoding in this context is that the main geocoding models for Brazilian addresses, such as the Google Maps (GOOGLE MAPS, 2025) and Mapbox (MAPBOX, 2025) APIs, are charged based on the number of requests and can generate high costs in big data processing, which is the case for most national governments. For instance, just three national government databases in Brazil—the driver's license, voter registration, and Individual Taxpayer databases—contain over 450 million address records. A second possibility would be to train or tune a custom geocoding model to improve the accuracy of Brazilian addresses; however, this process requires substantial training data and significant computational resources (ZHANG; BETHARD, 2025). Finally, note that most methods and models developed focus on geolocating addresses and calculating proximity between them, which may not accurately reflect the probability that two people know each other.

To overcome these limitations, this work proposes constructing a graph-based discovery model, grounded in the semantics of the postal code database, to calculate social proximity between two people using elements of their addresses. The idea is to expand the graph model proposed by (LIU *et al.*, 2023) for address matching to encompass the concept of address hierarchy and a weighted graph to propose a metric for the indicator of possible proximity between two people. This proximity index uses the Brazilian National Address Registry for Statistical Purposes and the hierarchical structure of the Brazilian National Postal Code Database. The hypothesis is that it is possible to discover a degree of familiarity between two people through graphs and address-related context information.

In this regard, the Brazilian postal code database was also essential for enabling a decrement

factor based on the distance between two addresses in the graph. The CNEF database enabled an individualized analysis of the probability that an individual knows another person within a given location, based on the total number of residences in each area. The Individual Taxpayer Registry in Brazil database enabled testing of the proposed model using addresses of Brazilian citizens. The initial modeling was performed to capture the semantics associated with postal code data. In this regard, entities and relationships were identified based on the different types of Postal codes as defined in Figure 7.2.

a) Degree of familiarity Index (DFI)

This basic strategy allows us to define the probability that two people who belong to the same Postal code know each other, but what if they belong to different Postal codes? This case is precisely where we leverage the power of graphs. The calculation involves the values associated with the vertices along the path between the two addresses, reducing the probability based on the types of entities encountered and the number of hops between them.

The familiarity index is a measure of the likelihood that two people know each other. Its calculation is based on two main premises: the possibility that two people at the exact location know each other depends on the number of people in that location (F1 - Local Population Density); the probability that two people from different regions know each other decreases substantially as the distance between them increases (F2 - Distance Decay Factor). As the index was designed, F1 effectively represents the initial percentage, and F2 is applied to this percentage as a decreasing probability factor.

The F1 calculation uses the CNEF database, which lists the residences in each zip code. According to the zip code database, not every higher-level location has a zip code, as do neighborhoods. In this case, the number of residences equals the sum of the home values in the lower-level locations. Another example of an area without a zip code is when the Thoroughfare is subdivided into sectors. In this case, the value will also be calculated based on the values of the hierarchically lower nodes.

The final structure of F1 is presented in the equation 7.1, where p_i is the population of the i-th subordinate location, n is the total number of subordinate locations, and P is the population of a location without subordinate locations. The final value of S can be the sum of the subordinate populations or the population of the location itself, if it has no subordinate

location. The final calculation of F1 is the minimum between 1 and 3/S. 3 was used as an approximation of the average number of people per household in Brazil, which in 2022 was 2.8 (O TEMPO, 2025), and the minimum to avoid index values greater than 1 in specific cases.

$$\text{F1} = \min\left(1, \frac{3}{S}\right) \quad \text{where} \quad S = \begin{cases} \sum_{i=1}^{N} p_i & \text{if } N \geq 1 \\ P_{\text{location}} & \text{if } N = 0 \end{cases} \tag{7.1}$$

Finally, it is essential to discuss distance-decay models (F2), since the level of relationship between two people is expected to decrease as they become physically farther apart. This phenomenon, known as the "Tobler's principle," states that all things are related to all others, but nearby things are more closely related than distant things.

One of the main proposed models, which has been fundamental in geography and urban planning for understanding how the interaction between two locations decreases as the distance between them increases, is the Inverse Power Law Model (PERLINE, 2005), presented in equation 7.2. In this equation, I_{ij} is the interaction between origin location i and destination location j, k is a proportionality constant representing the "strength" of attraction at the origin, d_{ij} is the distance between i and j, and β is the frictional exponent of distance. This parameter is the most critical, as it controls the decay rate. A high β (e.g., 2 or 3) means that distance has a powerful effect, and the interaction decreases dramatically with distance. A low β (e.g., 0.5 or 1) means the interaction is less distance-sensitive.

$$I_{ij} = \frac{k}{d_{ij}^{\beta}} \tag{7.2}$$

Since the distance in jumps in the graph significantly influences the likelihood of two people meeting, we will use a β equal to 3 in equation 7.2 to define the parameter F2, established in equation 7.3, where d_{ij}^3 represents the number of hops from node i to j, considering only the nodes that represent locations in the Postal code database.

$$\text{F2} = d_{ij}^3 \tag{7.3}$$

The final DFI formula is the application of the equation 7.3 to add the distance decrement value to F1, as presented in 7.4.

$$\text{DFI} = \frac{F1}{F2} \tag{7.4}$$

The DFI calculation does not apply to people living in the same residence because, in this case, the graph model assumes they know each other and that a relationship exists between them.

b) EXPERIMENT

To validate the process, we used the CNEF, Postal Code, and CPF databases, which were already modeled as graphs in Neo4J. We performed several searches using people's identifiers and names based on the CPF database. One of the searches performed is shown in Figure 7.15. Although we are dealing with open data, we have preserved people's identities and addresses, replacing them with Neo4j IDs.

Note that, according to the model, a person is linked to an address, which in turn is connected to a Postal Code. As defined in the indicator calculation, the number of hops considers only entities related to the location, so we begin calculating the indicator from the address. In the example presented, persons 3,204,350,038 and 376,162,656 are in the same Thoroughfare Sector, represented by the Postal code 17,800,100. In this case, the F1 calculation is limited to dividing the average number of families per household—3—by the number of addresses in the area—19. This number is then divided by 1 (hop) raised to the power 3, yielding a DFI with the same F1 value of 15.7895

The calculation between persons 237,869,189 and 2,982,291,059 requires the sum of all associated residences along the path between them. In this case, there are five residences in one sector of Rua Romao Martins and eleven in the other. It can be seen that the total sum of the Rua Romao Martins Thoroughfare is 16. The F1 calculation in this case would be 3/(16), or 18.75%. The calculation of the decrease factor takes into account the hop from the Thoroughfare sectors to the Thoroughfare, totaling two hops. Applying the formula presented in Equation 7.4, we would have a final DFI value of 2.3438%. Other realized experiments were documented in Table 7.5.

The experiment's results demonstrate that the semantic use of address data more meaningfully represents ties between two people than simply observing an association with the same

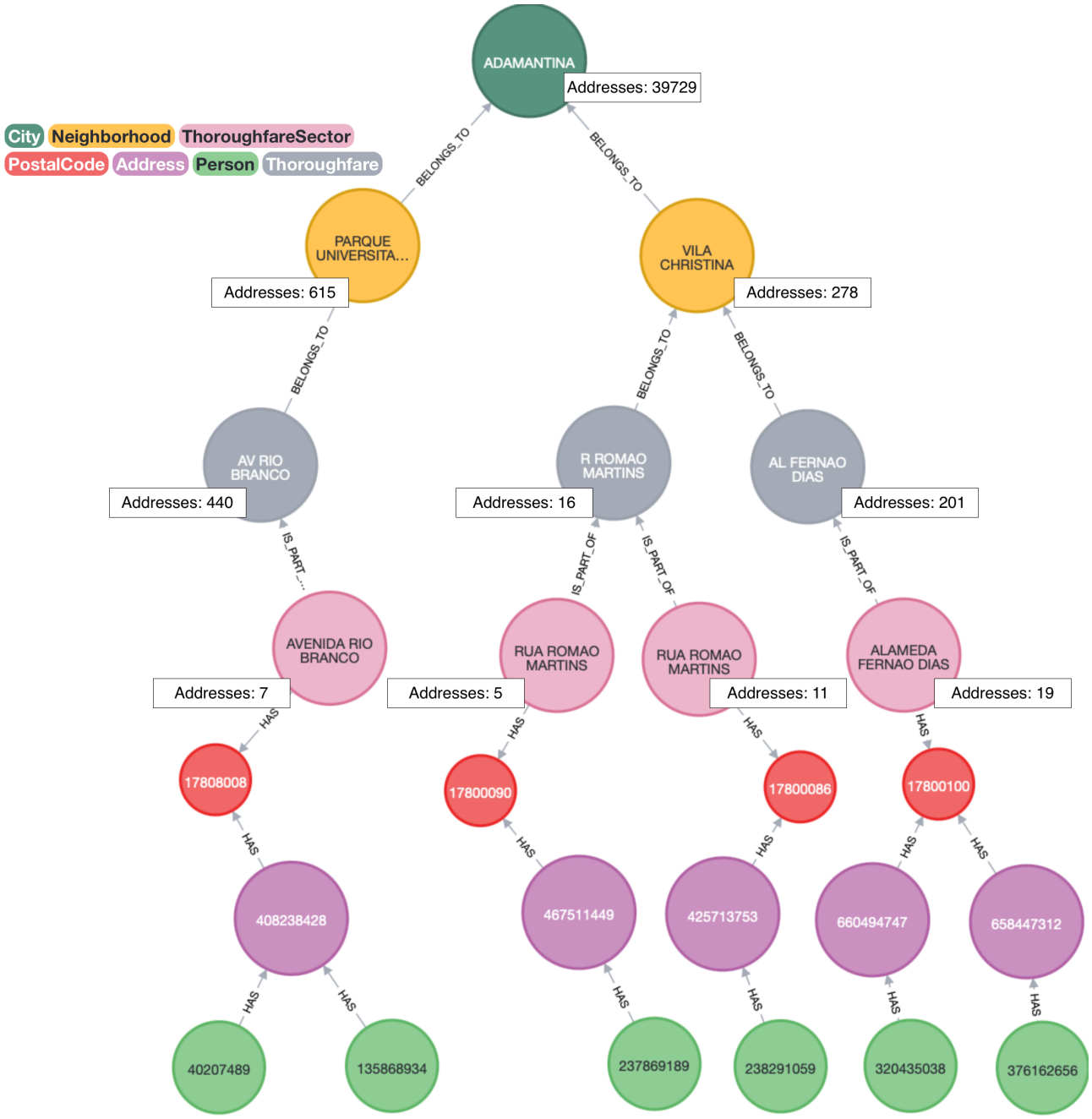

Figure 7.15: Path between people in Graph

Table 7.5: Experimental Results

Person/ People A	Person/ People B	Address Type	Dist.	Addr.	F1	DFI
40207489	135868934	Same Residence	-	-	-	-
320435038	376162656	Same Thoroughfare Sector - Al. Fernao Dias	1	19	15,7895%	15,7895%
237869189	298291059	Same Thoroughfare - Rua Romao Martins	2	16	18,7500%	2,3438%
320435038, 376162656	237869189, 298291059	Same Neighborhood - Vital Christina	3	278	1,0791%	0,0400%
40207489, 135868934	320435038, 376162656, 237869189, 298291059	Same City - Adamantina	4	39729	0,0076%	0,0001%

Postal code, assuming they are neighbors. This original measure proves insufficient, especially given the diversity of locations covered in the Brazilian postal database, which includes Postal codes with one and others with over 100,000 associated residences. Adding the semantics of the number of addresses proved to be an appropriate solution to address this problem.

On the other hand, it was also necessary to address cases where people had some relationship with each other, even if they were not associated with the same Postal code. Although in most cases the intersection of people from different regions yields a very low indicator value, in locations with a low number of residences, the indicator can identify a close relationship between two people. For future work, we suggest a more in-depth study of the threshold for determining whether a close relationship exists between two people.

From a scientific perspective, this analysis validates the application of graph topology as a proxy for physical proximity, extending Tobler's First Law of Geography (TOBLER, 1970) into the semantic space of knowledge graphs. Methodologically, the defined Degree of Familiarity Index demonstrates that exploiting the hierarchical semantics of postal codes offers a computationally efficient alternative to coordinate-based geocoding for social network analysis.

7.4.1.3 Non-trivial relationships between environmental transgressors

One of the main problems faced by AGU concerned the non-payment of environmental fines by violators of Brazilian ecological laws. The loss of revenue was approaching US$200 million in

2023 (TOLEDO, 2023). Most of the time, the real offenders, often large landowners, hid behind rural workers while carrying out illegal acts. This model's creation aimed to link offenders based on data from infraction occurrences, that is, the time frame and physical location (TORRES *et al.*, 2022). The premise behind this initiative is that these acts, such as illegal deforestation or land grabbing, often occur near the real offender's farm.

The open data on infractions listed in the previous section were processed to extract polygons of the affected areas. What was observed was an overlapping of polygons related to the locations of the infractions, indicating the criminal activity of different people in the same geographic region, as demonstrated in the red circles shown in Figure 7.16 (TORRES *et al.*, 2022). It was also observed that the activities of different people in nearby areas, usually in the same contiguous forest region, may also indicate correlated activities among individuals, presented in Figure 7.16, inside the yellow circle (TORRES *et al.*, 2022).

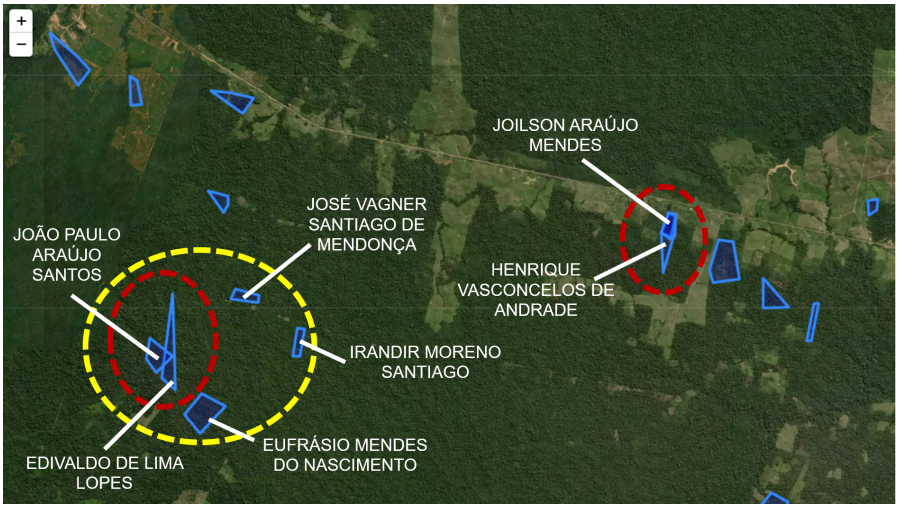

Figure 7.16: Correlation of events based on distance (TORRES *et al.*, 2022)

Density-based clustering methods (DB-SCAN) (ESTER *et al.*, 1996) were used to identify distinct groups/clusters in the data, based on the idea that a cluster in data space is a contiguous region of high point density, separated from other clusters by adjacent areas of low point density (TORRES *et al.*, 2022). The model evaluation was conducted in the state of Acre, Brazil. It involved 671 environmental occurrences with geometric points or polygons, creating 72 clusters, with 183 clustered and 488 isolated points, allowing the analysis of 618 individuals, since the same individual could be related to different assessments (TORRES *et al.*, 2022). A small subgroup of the clusters automatically generated by the tool can be seen in Figure 7.17

(TORRES *et al.*, 2022). In the figure, each group is identified by a color, and the red dots indicate isolated points, that is, places that cannot be grouped (TORRES *et al.*, 2022).

Figure 7.17: Groupings based on distance of events (TORRES *et al.*, 2022)

In practice, the model grouped 183 distinct records of environmental violations into 72 groups, each consisting of a minimum of 2 and a maximum of 6 violations (TORRES *et al.*, 2022). As one person or a group commits most infractions, some identified groups allowed the correlation to exceed a dozen people. As seen in subsequent sections, this joint analysis is the key to determining the real offender (TORRES *et al.*, 2022).

After creating the clusters, the network analysis and visualization began, generating non-trivial links to be imported into Neo4J. The use of an intermediate entity, a cluster, to carry out the interconnection of people, instead of performing the direct connection(TORRES *et al.*, 2022).

The final modeling proposal defines each "cluster" type entity as a location or geographic area where the offenses were committed (TORRES *et al.*, 2022). These areas are the key to correlating the "Person" type entities, representing the individuals effectively involved in the commission of the infractions. Part of the comments and relationships graph automatically generated by the model is shown in Figure 7.18 (TORRES *et al.*, 2022).

In the analysis of cluster number 22 (Figure 7.18), two more cluster-type entities, numbers 17 and 21, and another seven associated person-type entities were observed (TORRES *et al.*, 2022). The three infraction notices linked to cluster 17 occurred within a maximum interval

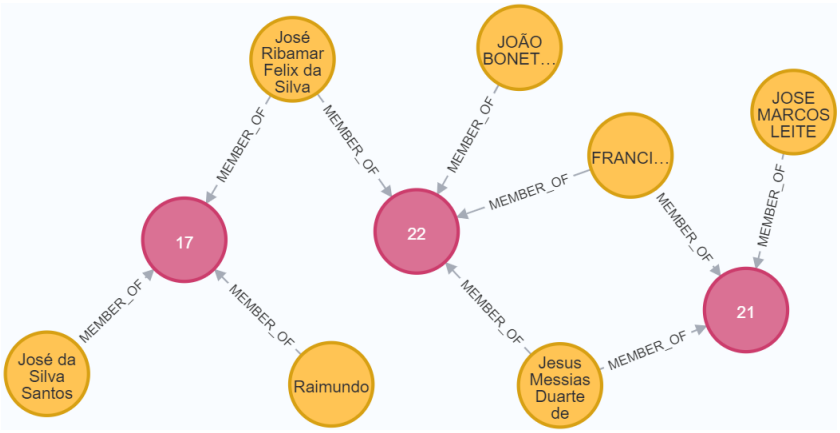

Figure 7.18: Graph of relationships between environmental transgressors (TORRES *et al.*, 2022)

of two days, between September 16 and 17, 2016. A similar time lapse is observed for a couple of the three associated notices to cluster 21, between September 3rd and 4th, 2016, and three of the four records associated with cluster number 22, between July 15th and 16th, 2014. Furthermore, from cluster number 22, it is also possible to detect a probable kinship relationship between two individuals who share the same surname-"FELIX" and "SANTOS" (TORRES *et al.*, 2022).

Concerning cluster 52 (Figure 7.19), the analysis of the records shows that 11 of the 12 occurrences were registered between the 6th and 10th of July 2017 (TORRES *et al.*, 2022). In addition, there was a coincidence of the surname "OLIVEIRA" in five of the 12 individuals, "SILVA" in four of the individuals, "LIMA" in two, and the surname "ANGLES" in another two, pointing to possible kinship relationships (TORRES *et al.*, 2022).

This experiment validates the scientific capacity to detect collusive behavior by spatiotemporally clustering seemingly independent events. Methodologically, it demonstrates the efficacy of integrating density-based clustering with graph projection, allowing the transformation of unsupervised learning outputs into explicit graph entities (clusters) that reveal hidden community structures among offenders.

7.4.2 Match Entities

The entity-matching processes in this experiment involved only direct cross-referencing using partial identifiers for people associated with the name. This was due to the types and origins of

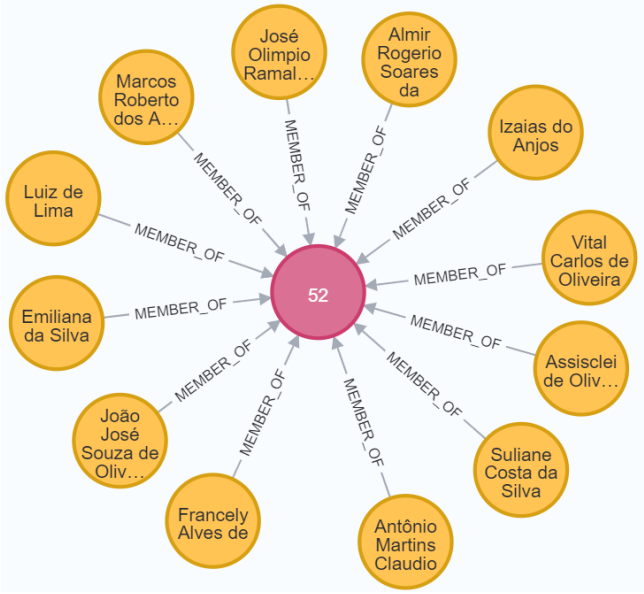

Figure 7.19: Graph of relationships between environmental transgressors

the data used in the experiment, which were consistent with the organization's maturity level. For this reason, it was not necessary to use more advanced entity-matching heuristics in this case.

7.4.3 Complete Knowledge Graph

Investigating asset concealment is, in essence, a task of network analysis. The strategies individuals and organizations employ to obscure asset ownership rarely involve a single step. Instead, they manifest as chains of interconnected relationships—family ties, corporate partnerships, indirect financial transactions, and the use of intermediaries. Representing these scenarios as a graph, with entities as nodes and relationships as edges, is an intuitive methodological approach. However, for the search for hidden assets to be effective, it is necessary to formalize and computationally execute the investigation rules and heuristics.

The objective of defining heuristics for asset concealment is not to find any arbitrarily complex path, but rather to identify specific, delimited, and semantically significant paths. An abstract query, such as "find any connection between a person and an asset," may generate results. However, these results would be difficult to interpret and defend as evidence of specific behavior. For this reason, it was necessary to formalize the rules for searching for these hidden assets.

Because the graph structure is represented in Neo4j, it was decided to adopt a declarative rule formalism, describing the structure of the suspicious situation using the standard Neo4j language, Cypher. In this paradigm, it is necessary to declare the pattern of interest, and the database system optimizes the search to find all instances that match that pattern in the data universe. It is also important to emphasize that the extensive use of Cypher in industry and research for fraud detection, anti-money laundering, and asset tracking also underpins this choice. In this sense, the formalization of the rules for identifying asset concealment is listed in Table 7.6.

Several points deserve emphasis regarding this rule-based analysis process. The first is that the rules presented in Table 7.6 must be applied in a complementary fashion; that is, rules governing relationships between people, such as R3, must be used together with rules associating people with assets, such as R1. It is also possible to combine relationship and asset association rules, such as R1, R4, and R5, allowing the location of assets in complex relationships, for example, a person who lives near another person who is a partner in a company that owns an asset. A visual representation of the rule composition for identifying hidden assets is presented in Figure 7.20.

Beyond their operational utility, the formalization of these heuristics constitutes a significant scientific contribution to the field of Knowledge Engineering. It represents the externalization phase of knowledge management, converting the tacit, experience-based intuition of investigators into explicit, machine-executable Cypher patterns. This formalization creates a reproducible, auditable ontology of fraud concealment schemes that can be shared, tested, and refined across different jurisdictions.

7.4.4 Deploy Knowledge Graph

At the end of the implementation process, a graph with 976,843,239 entities and 1,197,618,013 relationships was deployed in the production Neo4j Database. Table 7.7 details the element count for each of the 31 loaded entity types (labels), highlighting the scale and semantic diversity of the model, with over 253 million Asset entities and nearly 270 million Person entities. Table 7.8, in turn, presents the count for each of the 16 relationship types, illustrating the

Table 7.6: Graph Rules in Cypher

ID	Conceptual Heuristic	Formal Representation in Cypher Pattern
R1	**Direct Person Linkage**: An asset is held directly by a person.	`(p:Person)-[r:HAS]->(a:Asset)` `(p:Person)-[r:HOLDS]->(p:Position)-[r2:HAS]->(a:Asset)` `(p:Person)-[r:HOLDS]->(p:PensionRecord)-[r2:HAS]->(a:Asset)` `(p:Person)-[r:RECEIVES_RETIREMENT]->(re:Retirement)-[r2:HAS]->(a:Asset)`
R2	**Direct Organization Linkage** An asset is held directly by an organization.	`(o:Organization)-[r:HAS]->(a:Asset)`
R3	**Family Proximity** This heuristic tests the hypothesis of asset concealment through the transfer of assets to relatives (from first to fourth degree).	`path = (p:Person)-[:IS_MOTHER|IS_FATHER*1..4]-(p2:Person)`
R4	**Geographical Proximity** A person is linked to an asset held by an individual residing at a nearby address. Physical proximity may suggest collusion or the use of "straw men"who live nearby.	`(p:Person)-[r:HAS]->(a:Address)<-(p2:Person)` `(p:Person)-[r:HAS]->(a:Address)<-[r2:NEARBY]->(a2:Address)<-(p2:Person)`
R5	**Use of Companies** A person is linked to an asset througha chain of 1 to 5 relationships with companies. It uses a variable-length path in complex corporate structures. The depth is limited to 5 because a greater distance may indicate a tenuous connection	`path = (p1:Person)-[:IS_PARTNER]->(:Organization)-[:IS_PARTNER*0..4]->(:Organization)<-[:IS_PARTNER]-(p2:Person) WHERE p1 <> p2`
R6	**Shared Identifiers** This identifies hidden connections through shared contact information. Sharing a phone is a strong indicator of a close relationship, even in the absence of direct family or corporate ties.	`(p1:Person)-[r1:HAS]->(ph:PhoneNumber)<-[r2:HAS]-(p2:Person)` `(p1:Person)-[r1:HAS]->(ph:PhoneNumber)<-[r2:HAS]-(o:Organization)`
R7	**Joint Environmental Offenders** A person may be related to another environmental offender due to fines issued in nearby locations on close dates.	`(p1:Person)-[r1:IS_MEMBER]->(ph:EnvironmentalViolationCluster)<-[r2:IS_MEMBER]-(p2:Person)`

Figure 7.20: Graph Rules

density of connections established in the graph.

Table 7.7: Count of Elements by Entity (Label) of the Graph

Entity (Label)	**Number of Elements**
Address	234 527 339
AirCraft	15 968
Asset	253 496 467
Branch	61 778 218
City	5571
District	4304
EmailAddress	18 659 046
EnvironmentalViolationCluster	72
GovernmentDepartment	9330
LargeMailRecipient	20 250
MotorVehicle	95 599 515
Neighborhood	79 247
Organization	64 491 794
Partnership	437 215
Party	334 191 450
Payment	18 187 578
PensionRecord	131 033
Person	269 699 656
PhoneNumber	57 766 808
PhysicalPostOffices	12 742
Place	10 945
Position	74 156 003
PostOfficeBox	2164
PostalCode	1 504 735
Retirement	301 014
Salary	74 588 050
ShareCapital	64 491 794
Thoroughfare	1 281 192
ThoroughfareSector	193 515
Village	1070
WaterCraft	613 562

The implementation of Neo4j provides the first channel for interactive graph exploration: the Neo4j Browser, its native web interface. The result of a simple query performed in this tool is shown in Figure 7.22. In addition to interactive exploration, the Neo4j Browser interface enables data exploration via cipher using graph analysis algorithms. Finally, asynchronous APIs in Python were implemented to integrate with other systems. These APIs were integrated with some services available on the service portal for AGU employees.

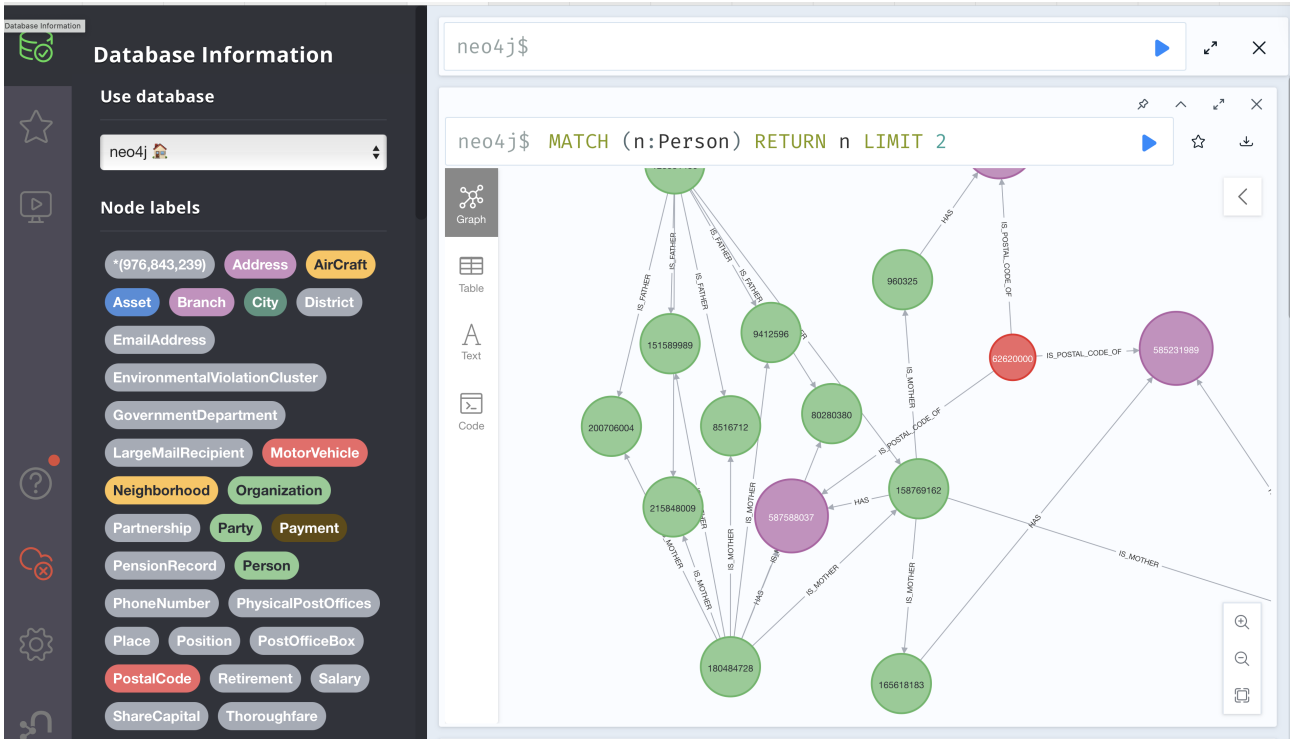

Figure 7.21: Neo4j Web Interface

Table 7.8: Count of Elements by Relationship Type

Relationship Type	Count
ACT_AS_REPRESENTATIVE	437 215
BELONGS_TO	1 499 785
HAS	608 435 960
HOLDS	74 287 036
INHERITS_BENEFIT_FROM	131 030
IS_FATHER	46 972 235
IS_MOTHER	108 035 015
IS_PARTNER	25 892 487
IS_PART_OF	193 515
IS_POSTAL_CODE_OF	238 343 087
MEMBER_OF	183
OF	437 218
OFFERS	74 457 017
PAYS	18 187 578
RECEIVES_RETIREMENT	301 014
REPORTS_TO	7638

7.5 KNOWLEDGE GRAPH USE AND EVALUATION

7.5.1 Visualize and Explore Knowledge Graph

One of the main ways to interact with and explore the Knowledge Graph was through Neo4j's native web interface, the Neo4j Browser. This tool allows executing Cypher queries and returns visual feedback on the results. This feedback can be provided through data tables or even by rendering the nodes and relationships found as an interactive subgraph. This allows the user to visualize the structure of connections, validate connection hypotheses, and intuitively understand the data topology, which is essential for analyzing complex networks.

The web interface also offers features such as Cypher syntax autocompletion, query history, and the ability to inspect the properties of each node or relationship with a simple click. Another positive point is its accessibility, as it does not require the installation of any additional software on the user's machine; a web browser is sufficient to connect to the database instance, allowing for rapid prototyping of search logic and the performance of ad-hoc exploratory analyses directly on the graph, as presented in Figure 7.22.

To facilitate the use of the Neo4j browser, queries based on the rules established in Table 7.6 were provided for execution in the web environment. The main query used is described in Listing 7.1 and searches for assets within up to 8 degrees of separation from the target individual.

```
-- 1. Identify the investigation target
MATCH (p1:Person {cpf: 'XXXXXXXXXXX'})

-- 2. Find a path from 1 to 8 steps ending in an asset,
--    using the complete and updated list of allowed relationships.
MATCH path = (p1)-[:IS_FATHER|IS_MOTHER|HAS|IS_PARTNER|NEARBY|IS_MEMBER
   *1..8]-(asset:Asset)

-- 3. Apply path validation filters
WHERE
  -- MAIN CONDITION: Ensures that ALL intermediate nodes
  -- belong to the complete list of allowed types in the diagram.
  ALL(n IN nodes(path)[1..-1] WHERE
    'Person' IN labels(n) OR
    'Address' IN labels(n) OR
    'PhoneNumber' IN labels(n) OR
    'Organization' IN labels(n) OR
    'Environmental Violation Cluster' IN labels(n) -- <-- ADDED
  )

  -- SECONDARY CONDITION: Ensures the final owner (penultimate node) is a
```

```
    Person
  AND 'Person' IN labels(nodes(path)[-2])

  -- SAFETY CONDITION: Ensures the final owner is not the investigated
    person themselves
  AND p1 <> nodes(path)[-2]

-- 4. Return the results for analysis
RETURN
    p1.name AS Investigated,
    nodes(path)[-2].name AS FinalOwner,
    length(path) AS SeparationLevel,
    labels(asset) AS AssetLabels,
    id(asset) AS AssetId,
    path
ORDER BY SeparationLevel
```

Listing 7.1: Cypher query to identify hidden assets through paths up to 8 degrees of separation.

The query in Listing 7.1 was executed in the Neo4j Web Interface for a set of debtors to the federal government, related to environmental infractions, obtained from the IBAMA database. An example of the query result for one of the debtors is presented in Figure 7.22. Note that the graph successfully identifies hidden assets at varying distances from the principal debtor, which may represent a form of asset concealment. Based on this information, it is up to the investigators to conduct a more in-depth analysis of the relationships identified to validate the practice. It is important to emphasize that the query only guides investigators within a vast universe of data, and human decision-making is essential to continue the investigative process.

7.5.2 Evaluate Knowledge Graph

The evaluation was based on two main, interconnected aspects that help validate the central hypothesis of this research. The first aspect refers to a qualitative assessment, developed to assess the investigative support utility of the graph, that is, its ability to answer complex questions and generate actionable intelligence for the lawyers of the Attorney General's Office. The second is the quantitative evaluation, which aims to examine the graph's structural integrity and the scale of data integration achieved, as well as to evaluate the performance of the models and AI developed in the knowledge enrichment section.

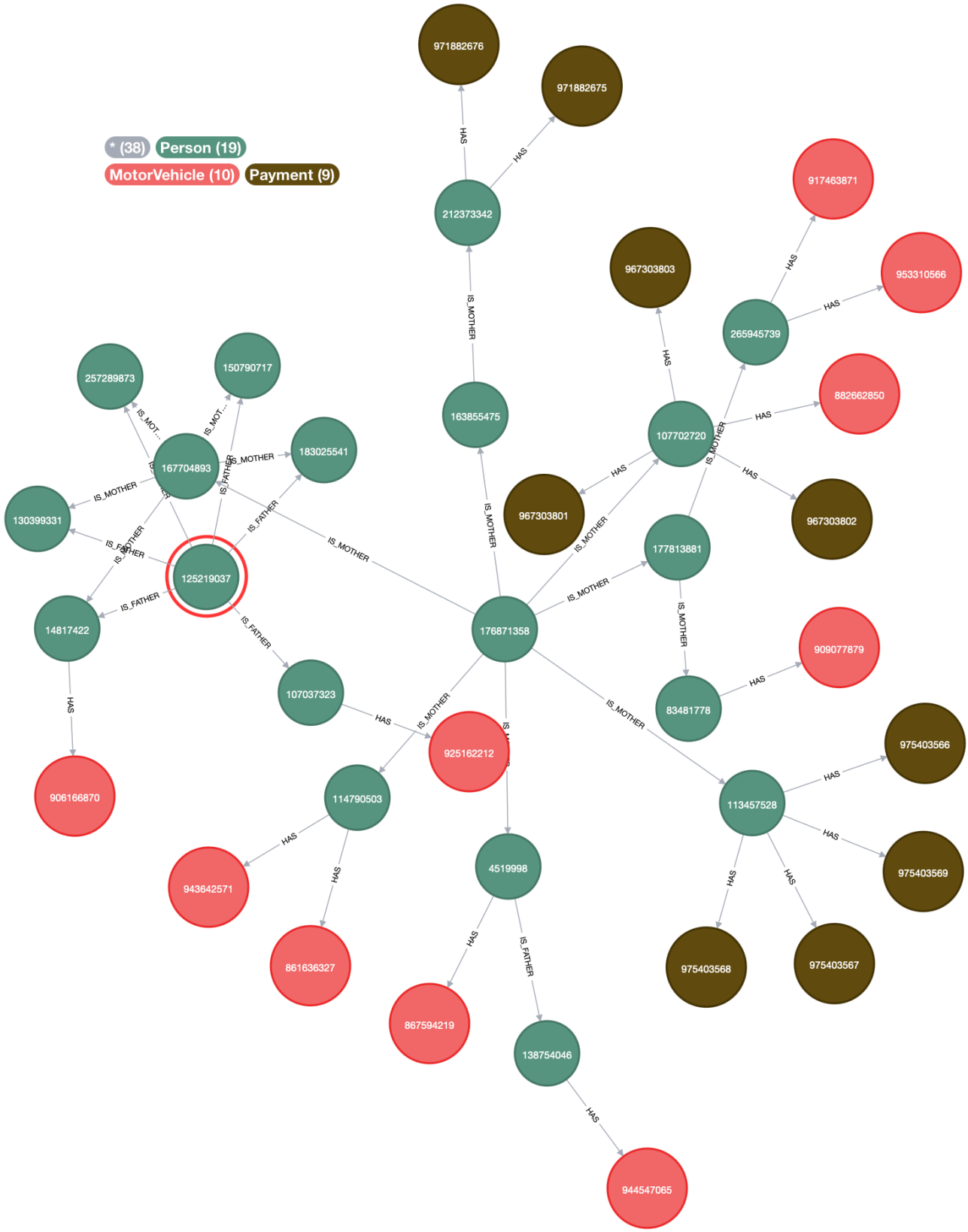

Figure 7.22: Asset Search

7.5.2.1 Qualitative Evaluation

This qualitative evaluation demonstrates how the KG, when queried with formalized heuristics, effectively discovers asset concealment patterns that would be extremely difficult or impossible to detect using traditional means. Table 7.6 formalizes the tacit knowledge of experienced federal attorneys in investigative patterns, in executable Cypher queries. The KG's ability to process these queries validates the completeness of its data model and the richness of its integrated connections.

Rules R3 (Family Proximity) and R4 (Geographic Proximity), also from table 7.6, test the hypothesis that assets are frequently hidden in the name of relatives or "front men" who reside nearby. The successful execution of these queries in the KG demonstrates that the integration of registration data, which contains affiliation and address information, was performed in a way that allows traversing multiple degrees of family separation and identifying neighborhoods, respectively.

Rule R5 (Use of Companies) is particularly essential, as it aims to detect the use of company chains to distance the debtor from the asset. The query tests the graph's ability to navigate complex corporate structures by executing a variable-length path pattern. Success in this task confirms that the CNPJ data, including the corporate relationships between individuals and legal entities, were effectively modeled and integrated.

The Shared Identifiers Heuristic (R6) explores a fundamental investigative insight: that sharing contact information, such as a phone number, is a strong indicator of a close relationship, even in the absence of formal family or business ties. The KG's ability to identify two people or a person and an organization connected by the same PhoneNumber node validates the entity resolution layer and the importance of modeling the main entities and their attributes as connecting nodes.

The execution of these heuristics on the graph materializes one of the main arguments of this thesis: that the ecosystem functions as a translator of human expertise, capturing the investigative logic that previously resided only in analysts' experience and transforming it into a formal, repeatable, and scalable computational process. An investigator's knowledge of "looking for relatives," for example, is translated into the Cypher query of rule R3, which in

turn is executed in seconds on millions of entities, a scale unattainable for manual analysis.

The query presented in Listing 7.1 is the synthesis of heuristics, designed to, starting from a research target (p1), discover assets within up to 8 degrees of separation, traversing only allowed relationship types and intermediate nodes. The results presented in 7.22 represent a concrete case study of the graph's effectiveness.

To formalize this qualitative assessment, we adopted the Competency Questions methodology (MONFARDINI *et al.*, 2023). The heuristics formalized in Table 7.6 serve as a set of validation scenarios, ensuring that the graph's semantic structure can answer complex investigative inquiries—such as identifying indirect ownership or geographic collusion—that were previously impossible with fragmented data sources.

7.5.2.2 Quantitative Evaluation

The validity of the Knowledge Graph is supported not only by its qualitative utility but also by its scale, the integrity of its structure, and the performance of the artificial intelligence components used in its construction. Regarding Scale and Topology Analysis, the final metrics indicate that more than 2 billion entities and relationships were extracted from 15 datasets (Tables 7.7, 7.8). These numbers measure the success of the ecosystem's main proposal to overcome information fragmentation. The ability to integrate and unify almost one billion entities from dozens of data sources is concrete proof of the architecture's and the implemented data pipelines' technical viability and scalability.

The diversity of labels across 31 entity classes highlights the semantic richness of the model. It successfully represents the heterogeneous network derived from the domain's complexity and integrates biographical, corporate, asset, geographic, and employment data into a single cohesive conceptual scheme.

Regarding evaluating the Relationship Discovery Models, the kinship discovery model achieved 97% accuracy for the TRUE class (valid kinship), with a recall of 41

The DFI evaluation, however, is not based on accuracy/recall metrics, but on its innovation and pragmatism. By leveraging the hierarchical semantics of the Brazilian postal code system and the number of residences per location, the model can infer a proximity index without

relying on geocoding APIs, which would be financially and computationally prohibitive for the scale of the project's data. The main contribution of this model is its efficiency and domain-knowledge-based approach, which transforms a simple textual attribute (the postal code) into a rich source for proximity inference.

Regarding query time performance, some query executions with up to 8 levels of depth, such as the one shown in Listing 7.1, were performed. The Table 7.9 presents the result of executing a sample with five examples, including the searched CPF (Brazilian taxpayer ID), the number of nodes returned, and the query time, and it demonstrates excellent performance.

From a formal quantitative perspective, the system's performance was evaluated using standard information retrieval metrics. The Kinship Discovery model, for instance, prioritized Precision (97%) over Recall (41%) to minimize false positives, a critical requirement in legal investigations. Furthermore, the graph's structural analysis confirms the architecture's scalability, successfully integrating over 2 billion elements while maintaining millisecond latency for critical relationship queries.

Table 7.9: Execution Results of the Hidden Asset Search Heuristic

Target CPF	**Nodes Returned**	**Total Time (ms)**
1********34	1408	104
6********63	2882	177
0********08	62	15
1********87	54	14
1********87	2343	186

7.6 SUPPORT ACTIVITIES

7.6.1 Ontology Improvement

The development and governance teams treated an Asset Concealment Ontology as an element in constant refinement, aligned with the methodology proposed in Chapter 4, which advocates a data-driven approach, and with the support activity defined in Chapter 5. The evolution process used the systematic application of the predicted feedback loop, in which each data integration challenge became a requirement for refining the semantic model. Multiple

change requests were submitted to the knowledge management team and governance committee, most of which were approved, resulting in significant changes to the reference ontology throughout the process.

The first request concerned the initial data analysis from the CNEF and the CEP. This analysis revealed that a flat address model was insufficient for the problem. The hierarchical structure of the Brazilian postal code contained a semantic that could be explored to infer geographic and, hypothetically, social proximity, as explored in Section 7.4.1.2. Refining the ontology to mirror this hierarchy and capture this semantics was necessary. Because of that, new classes were introduced, such as ThoroughfareSector, District, and Village, and relationships between them were established, as illustrated in Figure 7.2.

This decision to transform a simple text attribute (the postal code) into a navigable structure in the graph was made under the supervision of the knowledge management team and approval of the governance committee. The direct consequence was enabling proximity queries without needing geocoding, a computationally and financially costly solution.

Another modeling challenge arose with data from RAIS and CAGED. These presented a classic problem: an n-ary relationship. The employment relationship connects a Person, a Company, and a Salary (an Asset). A simple binary relationship 'works_in' between Person and Company and 'has' between Person and Salary would miss the crucial connection of which job corresponds to which salary, especially for individuals with multiple employment relationships.

To solve this, we propose that the knowledge team adopt the ontological design pattern known as reification (n-ary relation pattern), creating a new class, Position, that represents the employment relationship itself. This new class was then connected by binary relationships to the three original entities: (Person)-[holds]->(Position), (Position)-[offer]<-(Organization), and (Position)-[has]->(Salary), as detailed in Figure 7.6.

Data analysis from CNPJ revealed that the 'is_partner' relationship was sometimes mediated by a third entity: the legal representative of a partner class. This class was modeled to represent the shareholding itself, allowing us to connect the shareholder company and its legal representative to that specific shareholding, as outlined in Figure 7.5. This seemingly minor adjustment is crucial for the legal defensibility of the graph's inferences. In an investigative context, the distinction between being a shareholder and a shareholder's representative is le-

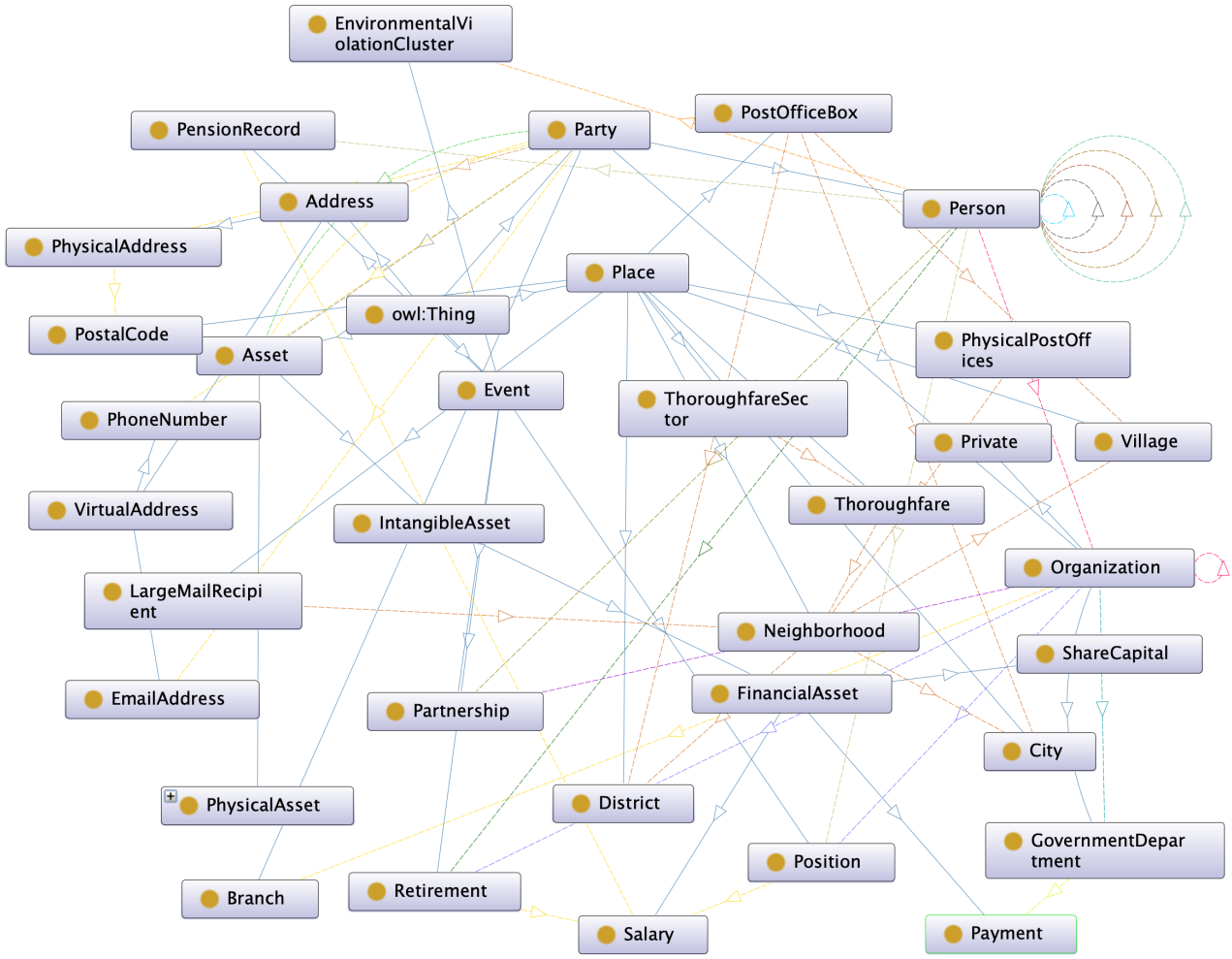

Figure 7.23: Final Ontology

gally fundamental. In this case, the ontology improvement was directly driven by the need for legal precision.

These change requests were the ones that led to the most significant alteration of the reference ontology. Combined with other minor requests, they led to the definition of the final ontology presented in Figure 7.23.

7.6.2 Data Governance

One of the biggest data quality challenges was the inconsistency of identifiers. For privacy reasons, many open datasets presented the Brazilian taxpayer ID (CPF) of partners in a partial (masked) form. Furthermore, the problem of homonyms was endemic, especially in affiliation fields. The solution was a governed process. First, the CPF was established as the master

identifier for the Person entity, a Master Data Management decision, and a key governance component. Second, for partial CPFs, the Data Stewards team developed and validated a deduplication heuristic that cross-referenced masked CPFs with full names, as described in Section 7.3.1.

For the homonym problem in affiliation, the solution was even more sophisticated. A machine Learning model (Section 7.4.1.1) was developed, with its construction and validation supervised by AGU lawyers to ensure the accuracy and legal relevance of the inferred links. Data governance in this project manifested itself very much as an active process of resolving quality problems, combining business rules, heuristics, and, when necessary, AI models, demonstrating a mature approach that treats data quality as a continuous curation process.

Furthermore, the project handled a massive volume of sensitive personal data in compliance with the General Data Protection Law (LGPD), which requires purpose, necessity, and security in data processing. Data governance was the mechanism we used to ensure compliance. Ingestion pipelines were designed to load only the attributes strictly necessary for analysis (principle of necessity). Access to the graph was segmented by profile, ensuring that only authorized users could access sensitive information.

7.7 DISCUSSION AND CRITICAL ANALYSIS

The implementation of this ecosystem raises critical points regarding the human role, ethics, and auditability in automated investigations. First, the system operates under a "human-in-the-loop" paradigm; the Knowledge Graph serves as a decision-support tool, enhancing the investigator's cognitive capacity rather than replacing their judgment. The graph identifies potential concealment paths, but the legal qualification of fraud remains a human prerogative.

Ethically, while the graph integrates vast amounts of data, it adheres to the principles of purpose and necessity mandated by the GDPR/LGPD. However, aggregating data creates a mosaic effect, potentially revealing sensitive inferences from non-sensitive data, which requires strict access governance. Regarding auditability, the use of a declarative graph query language offers superior transparency compared to black-box deep learning models, allowing every inference to be traced back to its source facts and relationships, ensuring the legal defensibility of

the findings.

Finally, we must address operational risks and limitations. The "Garbage In, Garbage Out" principle remains a significant risk; errors in source databases (e.g., outdated addresses) are propagated to the graph. Additionally, maintaining the ontology requires continuous effort to prevent semantic obsolescence as fraud schemes evolve. The reliance on deterministic rules, while explainable, may also limit the system's ability to detect novel, previously unknown fraud patterns that do not fit the pre-programmed heuristics.

7.8 SUMMARY

This chapter detailed the ecosystem's practical implementation and validation in a real-world environment, applying the Knowledge Graph construction methodology (defined in Chapter 5) and using the Asset Hiding Ontology (developed in Chapter 6) as a semantic schema, the ecosystem successfully integrated 15 heterogeneous, large-scale government data sources, including CPF, CNPJ, RENACH, and RAIS, thereby overcoming the architectural fragmentation described in previous chapters.

The success of this implementation depended on the innovations proposed by the KG construction process, which addressed the methodological fragmentation. Among the results achieved, in addition to the instantiation of the knowledge graph for asset hiding, we can mention the development of AI models and customized heuristics—notably for Discover Kinship and Social Proximity—essential for unifying identities and discovering non-trivial relationships. The next chapter will consolidate and discuss the formal results of this implementation.

CHAPTER 8

RESULTS AND DISCUSSION

This chapter discusses and presents the most significant results of this work. It involves proposing methodologies for developing ontologies and constructing knowledge graphs, developing an ontology for the domain of asset hiding, and implementing and validating the proposed ecosystem.

8.1 PROPOSAL OF AN ONTOLOGY BUILDING PROCESS RESULTS

The process proposed for developing the ontologies used in this research can serve as a template for future projects aiming to create domain-specific ontologies, especially in areas that require integrating extensive, heterogeneous data sets (TORRES *et al.*, 2024).

In practice, this process stands out from other methodologies for several vital reasons. Unlike some methods that focus on conceptualization without much attention to real-world data integration, this process strongly emphasizes data integration from the outset, including internal and external datasets (TORRES *et al.*, 2024). Additionally, it points to improvement in essential aspects not addressed in these original methods, particularly concerning its scoping strategy, ontology reuse and alignment, and ontology validation.

8.1.1 Ontology Domain Scope

The first significant improvement concerns the definition of the application domain and the scope of the ontology. One of the main difficulties in the ontology modeling process and the development of systems is precisely the gathering of requirements. This problem is common in ontology modeling, where the challenge is identifying datasets relevant to the business being modeled.

In this sense, the proposal defines a strategy based on predefined categories of interest to

narrow the search scope and facilitate the identification of datasets relevant to domain mapping. At the same time, the proposed model extends the search beyond the organization's boundaries, enabling the identification of datasets related to the domain but outside the organization. This fact allows the ontology to be built with a focus on interoperability among the various actors operating in the domain, facilitating the achievement of its primary objective of providing interoperability without requiring significant changes to the scope of concepts and relationships identified.

8.1.2 Ontology Alignment, Integration, and Reuse

The second point of improvement concerns the integration and use of all or part of existing ontologies. (USCHOLD; GRUNINGER, 1996) also characterizes it as a complex problem. The proposed idea was to systematize a process that was highly conceptually constructed. The analyzed processes make it clear that it is necessary to use existing ontologies, even if only partially, to improve integration. However, an effective strategy for carrying out such an operation has not been defined. In this context, we proposed a method for aligning ontologies during their construction phase to address this problem.

8.1.3 Ontology Evaluation

The proposed method goes beyond formal ontology validation. Also, it assesses its practical applicability by requiring testing in a real project, thereby reducing the probability of changes to the ontology's concepts and relationships when applied in production. This action is beneficial because it reduces potential impacts on the development of applications that use the ontologies, as development teams would need to wait for the adjustments before continuing their activities. In addition, it avoids the need to maintain a team of knowledge engineers throughout the application development project's lifecycle that uses ontology, thereby reducing the project's total costs.

This proposal was conceived based on two main aspects. It was observed that including the ontology construction process within the scope of knowledge graph construction has both

negative and positive aspects. The negative element stems from the need to simplify a naturally complex process so it fits as a step in a broader graph-construction process. Furthermore, during the construction of the Asset Concealment graph, which will be described in later chapters, specialized knowledge-engineering teams were allocated throughout the process, often underutilized, corroborating the need to treat ontology construction separately.

However, on the positive side, linking the ontology construction to a vision of practical application directs its use towards something that generates value and is applicable. From practical experience in projects, it is not uncommon for an ontology to be developed without achieving its potential for use within organizations. It is because, as traditional ontology creation processes are separate from application creation processes, communication failures often prevent needs from being adequately met by different teams.

In this way, the proposed model maintains independent ontology construction but includes a series of validation steps directly related to using ontologies in real applications. This initiative is a middle ground between ontology construction and knowledge graph processes, keeping the ontology construction process separate while establishing a strong link between the theoretical model and practical application, thereby enabling validation and enhancing the future use of the planned structures.

8.1.4 Table Summary

The main contributions of this work regarding the ontology definition process are presented in Table 8.1.

8.2 PROPOSAL FOR A KNOWLEDGE GRAPH BUILDING PROCESS RESULTS

The new process proposed offered significant benefits in terms of organization, role clarity, and execution efficiency. One of the highlights is the model's iterative nature, which allows revisiting previous steps whenever additional business requirements, demands to update the ontology, or discoveries of relevant data are identified. It reduces the rework and enables specific adjustments as new opportunities or challenges emerge.

Table 8.1: Comparative Analysis of Ontology Building Processes

Dimension	Proposed Process (Chapter 4)	Traditional Methodologies (Section 2.2.1.3)	Comparative Analysis
Main Focus / Starting Point	**Data-Driven**: The process begins with "Data Cluster Identification", analyzing real-world data sources to derive concepts.	**Generally Conceptual (Top-Down)**: Many methodologies start from informal "competency questions" or expert knowledge elicitation to define abstract concepts.	The proposed approach innovates through its pragmatism, anchoring the ontology in the reality of available data from the very first step and ensuring the model is directly applicable and useful for integrating government data sources, thereby mitigating the risk of creating a theoretical schema disconnected from investigative practice.
Evaluation Strategy	**Multifaceted and Application-Oriented**: Proposes a robust, three-pronged evaluation: Data Sample Mapping, User-based Evaluation, and Application-based Evaluation.	**Focused on Logical Consistency and Conceptual Coverage**: Traditional evaluation often focuses on verifying logical consistency (using reasoners), answering formal competency questions, and expert review.	The inclusion of "Application-based Evaluation" and "Data Sample Mapping" is a critical differentiator. It tests not only whether the ontology is theoretically *correct*, but also whether it is practically *useful* and *complete* when instantiated with real data and integrated into a system (the Knowledge Graph).
Reuse and Alignment	**Formal and Structured Step**: Dedicates two distinct activities: "Reusability Review" and "Existing Ontology Alignment", using systematic tables for mapping.	**Recommended, but with Less Formal Structure in the Process**: Methodologies like "Ontology Development 101" and Methontology recommend reuse. Still, the alignment process is not detailed with the same level of methodological formalism.	The proposal elevates alignment from a best practice to a formal and explicit methodological step. The use of tables to map concepts and relations makes the integration process more rigorous, auditable, and reproducible.
Expert Involvement	**Continuous and Structured Collaboration**: Domain experts participate from "Scope Definition" to "User-based Evaluation", ensuring constant alignment with investigative needs.	**Focused on Specific Phases**: Expert involvement is essential but often concentrated in the initial knowledge acquisition phases (e.g., interviews, competency questions) and final validation.	By integrating experts at multiple checkpoints throughout the cycle, the proposed process adopts a participatory design approach. It ensures the ontology not only represents the domain correctly but also does so in a way that is intuitive and useful for the analysts who will use it.

Another significant gain was the separation of the ontology process. Treating ontological modeling in a separate moment led to a more effective use of ontology experts, who could focus on the model's conceptual evolution.

The establishment of clear policies provided greater traceability and reliability of information

regarding data governance and quality. This point proved crucial in domains where legal compliance (e.g., GDPR) plays a central role, as adopting robust governance practices reduced the risk of misuse or duplication. In addition, continuous data quality verification and ontology evaluation helped to mitigate inconsistencies and ensured that the graph evolved in line with business needs.

Another critical point is the more effective division of roles and responsibilities among business experts, data engineers, ontologists, and developers. It became clearer when each profile should act, which mitigated the overload on specific teams and promoted better alignment among the different areas involved, favoring an agile, assertive workflow.

The authors of the original process also listed some points in their work that were addressed in this new model. The proposed model emphasizes user and business expert participation at different points in the flow, especially during the Business Understanding, Visualize and Explore Knowledge Graph, and Evaluate Knowledge Graph stages. In this way, the actual use of the graph, the needs of those who consult it, and the opportunities for improvement identified by users begin to directly influence the development and evolution of the graph — something that the literature review indicated as little addressed in processes strictly focused on algorithms or technical applications. Additionally, the proposed method was tested and validated in an extensive, complex case study involving multiple databases and people.

Table 8.2 presents a summary of the main improvements proposed in this work compared to the previous process defined in (TAMAŠAUSKAITĖ; GROTH, 2023).

8.3 PROPOSAL OF AN ASSET CONCEALMENT ONTOLOGY RESULTS

The ontology developed to investigate asset concealment demonstrated its effectiveness through successful implementation and validation in a real-world scenario involving a Brazilian public agency (TORRES *et al.*, 2024). Its integration into corporate systems significantly improved the detection and analysis of concealed assets, uncovering complex ownership structures and hidden relationships that were previously challenging to identify. During the Data Sample Mapping and Ingestion phase, real-world data samples were incorporated into the ontology to assess its structure and performance. The ontology effectively captured all critical concepts,

attributes, and relationships necessary for asset investigations, ensuring its adaptability to the complete domain-specific requirements.

In the User-based Evaluation, domain experts and stakeholders provided helpful feedback confirming that the ontology covered all pertinent concepts and relationships. Based on the suggestions, refinements that further changed the ontology improved its performance in investigative tasks (TORRES *et al.*, 2024). The ontology was validated for practical applicability in application-based evaluation in real-world work systems. When aligned as a knowledge graph, it could integrate well into existing processes and investigative systems and showed no logical inconsistencies. The ontology performed well by maintaining data integrity while simultaneously communicating with disparate sources and systems. The stakeholder feedback was incorporated continuously, enhancing the ontology's effectiveness and usability.

The methodology used to develop this ontology gives a structured framework for future projects requiring domain-specific ontologies, especially when working with large and heterogeneous datasets (TORRES *et al.*, 2024). More specifically, it stands out from orthodox methodologies in several crucial ways. Far from methods that emphasize theoretical conceptualization with little integration of real-world data, this process prioritizes the early and seamless incorporation of data from internal and external sources. It is deplorable in the context of asset concealment investigations, where it is critical to draw on disparate databases. The proposed methodology, in addition, talks directly about an issue that usually bedevils data integration: data silos, fragmented, misplaced, or abandoned knowledge that is splintered across different databases or organizations. Most formal ontology development approaches do not address this solution directly, hence failing in the initial domain-enlargement phase.

Most importantly, another innovative feature of the method is presented here: the Data Cluster Identification step, which clusters data by thematic relevance (i.e., persons, companies, and assets). Finding relevant data will be broader in scope, covering even sources that might fall outside an immediate investigation. Another big thing about the methodology is that it will include a Domain Knowledge Capture section that charts internal and external datasets relevant to the investigation, ensuring that ontology construction is directed toward practical application rather than purely theoretical.

The proposed methodology is designed for versatility, with applications beyond asset conce-

alment that include various investigative domains, such as fraud detection, money laundering, and cybercrime. Competent pattern recognition and data integration are key. Modular in design, the proposed methodology allows easy adaptation to various investigative areas; it will therefore be superior in scalability and reusability to other conventional methods, which focus on relatively minor, case-specific domains.

Using a hands-on approach to real-world data mapping has enabled the ontology to remain scalable while accommodating vast, diverse datasets from various sources, especially in disciplines where the volume and variety of data pose analytical challenges. Unlike classical ontology development procedures that often treat domain-specific integration superficially, this process is particularly suitable for large-scale, data-intensive applications because of its meticulous attention to integrating domain-specific data.

Another cornerstone of innovation for this methodology is unwinding alignment considerations, in which existing upper-level and domain-specific ontologies are integrated and aligned. Many traditional methods overlook the importance of alignment, sometimes treating it superficially. In contrast, the approach proposed herein systematically resolves inconsistencies and redundancies across multiple ontologies to ensure consistency and interoperability. Besides the intended reuse of existing ontologies, this process extends the ontology to a broader scope while minimizing redundancy. It uses an alignment strategy to harmonize overlapping concepts, making the ontology immediately practical and effective for investigative applications.

8.4 ECOSSYSTEM IMPLEMENTATION RESULTS

The fundamental result of this thesis is the successful unification of vast and heterogeneous government datasets into a single, cohesive knowledge graph. The scale of this integration constitutes the first and most direct proof that the methodological process proposed in Chapters 4 and 5 is scalable and effective in handling the complexity of the real world and more than 2 billion elements, including entities and relationships. Furthermore, to the best of our knowledge, we have not found a specific knowledge graph for handling asset concealment fraud in the literature, which makes the graph a significant finding of this research.

Additionally, integration at such a scale would have been impractical without the Asset

Concealment Domain Ontology acting as a unifying schema. The ontology provided the necessary semantics to effect the connection and demonstrated that a strong semantic foundation is not a byproduct, but a prerequisite for large-scale knowledge integration. The methodological validation of the knowledge graph construction process should be highlighted, particularly given its application in a real-world environment with a high volume of data, which goes beyond the academic scope of previously defined models. All the activities developed within the multiple stages of the process, from understanding the business, planning the deployment of the environment, acquiring and improving knowledge, using and evaluating the KG, and support activities, validated the proposed structure. It is essential to emphasize the addition of important points to the process, which differentiate it from academic models and allow its use in a real-world environment, such as support activities, like data governance, planning, and deploying the technological environment.

As described in the KG evaluation section in Chapter 7, the AI models built to enrich knowledge are also presented as essential results of this thesis. To the best of our knowledge, no model for automatically discovering hidden kinship links from data with homonyms has been found in the literature. The model developed in this work achieved 97

Another proposed model was the social proximity calculation, which was proposed to define a proximity indicator between neighbors or residents of nearby areas without geolocation data, but whose costs would be prohibitive for large-scale use and for the volume of government data. The proposed Degree of Familiarity Index uses the hierarchical structure of the Brazilian Address Code system and data from the National Census of Addresses for Statistical Purposes to calculate a proximity metric. The experimental results, presented in Table 7.5, demonstrated that the DFI varies logically and consistently with the semantic distance between addresses, which shows that deep and specific domain knowledge (in this case, the structure of the postal system) can be encoded in a model to produce a more subtle and contextually relevant metric than generic physical distance.

The last model presented was based on the DB-SCAN clustering algorithm and data from IBAMA's environmental infraction reports. It is still a preliminary study, but it appears to be positively biased. The analysis of 671 occurrences in the state of Acre identified 72 distinct clusters of people.

In the complete knowledge Graph section, it is worth highlighting the formalization of investigative tacit knowledge in the area of assets as a set of investigative heuristics, coded as Cypher queries (Table 7.6). The KG's ability to process these queries validates the completeness of its data model and the richness of its integrated connections. The consolidated query, presented in Listing 7.1, was designed to, starting from an investigation target, discover assets with up to 8 degrees of separation, traversing only permitted relationship types and intermediate nodes. Executing this query for real debtors related to environmental infractions produced encouraging results, as shown in Figure 7.22. The analysis of this example graph reveals a path of asset concealment with multiple degrees of separation, which would be virtually impossible to trace manually or through queries in isolated databases. The identified path connected the target debtor to an asset (in this case, a salary or receipt of public funds) through a chain of relationships.

8.5 SUMMARY

This chapter discussed and analyzed the research results, validating the thesis's central contributions. We structured the analysis to evaluate each proposal, demonstrating the effectiveness of the ontology construction methodology and the Knowledge Graph construction process. We also validated the completeness of the ontology for the asset concealment domain and the performance of the ecosystem implementation. The ability to integrate data and discover hidden relationships also validated the framework.

Collectively, the results confirm the central hypothesis of this thesis. The practical evaluation in AGU demonstrated that the proposed ecosystem is a viable and effective solution to overcome semantic, methodological, and architectural fragmentation. The final chapter of this thesis will consolidate these contributions and discuss future perspectives.

Table 8.2: Comparative Analysis of KG Building Processes

Dimension	Proposed Process (Chapter 5)	(TAMAŠAUSKAITĖ; GROTH, 2023)
Business Integration	**Dedicated and initial phase:** Proposed methodology begins with "Business Understanding", ensuring that all development is guided by business objectives, requirements, and success metrics defined in collaboration with domain experts.	**Implicit and Ad-hoc:** The participation of business experts is not formalized as a structured phase. It is implied in the "Identify Data" step, but lacks a defined process for continuous collaboration.
Data Governance	**Continuous support activity and central pillar:** Governance is planned from the outset ("Planning Data Governance") and maintained as a cross-cutting support activity ("Data Governance"), ensuring quality, compliance, and security throughout the entire lifecycle.	**Not explicitly addressed:** The academic framework focuses on the technical steps of graph construction but does not formalize data governance processes such as versioning, lineage, or compliance with regulations (e.g., GDPR), a critical gap for enterprise environments.
Ontology Lifecycle	**Decoupled and iterative process:** Ontology construction is treated as a separate process (Chapter 4), and its evolution is managed continuously through the "Ontology Improvement" support activity. It ensures a stable and adaptable semantic foundation.	**Linear and integrated step:** Ontology construction is treated as a single step ("Construct the KG Ontology") within the main workflow. It does not reflect the complexity and iterative nature of knowledge modeling in complex domains.
Infrastructure Planning	**Explicit and pragmatic phase:** Includes the "Environment Planning and Deployment" step, which formalizes the definition of the technological architecture and the deployment of the environment, bridging the gap between theoretical design and practical implementation.	**Not explicitly addressed:** The process does not include a formal step for planning the technological infrastructure, an essential step for executing real-world projects, as noted by the authors themselves as a limitation.
Iteration Model	**Agile and feedback-driven lifecycle:** The process is inherently iterative, with continuous support activities and the possibility of returning to previous phases (e.g., from evaluation back to business understanding).	**Predominantly linear:** The process is presented as a sequence of six main phases. Although it can be adapted, its fundamental structure is not designed as a continuous lifecycle but rather as a development pipeline.
Evaluation Method	**Focus on business value and technical quality:** The "KG Use and Evaluation" phase directly links the evaluation to the business objectives defined in the first step, using qualitative and quantitative metrics.	**High-level and general:** Evaluation is described generically, without detailing metrics or the distinction between the internal quality of the graph and its utility in practical tasks.

CHAPTER 9

CONCLUSION

The central problem this research aimed to solve is the State's ineffectiveness in combating asset concealment, a deficiency stemming from the profound semantic, methodological, and architectural fragmentation of its vast data repositories. This operational gap grants fraudsters a strategic advantage and limits the ability to recover public funds, perpetuating a cycle of impunity and the loss of essential resources.

Given this scenario, the following research question was formulated: How can the systemic and semantic integration of fragmented government databases enhance the Attorney General's Office's analytical capacity to detect asset concealment fraud proactively? To answer this question, the thesis was guided by the central hypothesis that creating a computational ecosystem, architected around a Knowledge Graph and guided by a domain ontology, could overcome this fragmentation, significantly increasing the State's capacity to detect and investigate these complex financial crimes.

The proposed solution materialized in a complete socio-technical framework, composed of: (i) innovative and enterprise-ready methodologies for the construction of ontologies and Knowledge Graphs; (ii) a pioneering domain ontology for asset concealment; and (iii) an end-to-end computational ecosystem, validated in a real-world environment in collaboration with the Attorney General's Office. The approach adopted recognizes that the solution to data fragmentation in high-consequence legal domains is not purely technological but socio-technical. The success of the ecosystem depends on the deep integration between technology (the Knowledge Graph), formal processes (the methodologies of Chapters 4 and 5), and human expertise (the continuous involvement of lawyers and analysts from the Attorney General's Office). The graph construction process, for example, explicitly begins with the Business Understanding and incorporates User-Based Evaluation as a central validation step, demonstrating a design philosophy where technology is built to augment, not replace, the human investigator.

The central hypothesis of this thesis was empirically validated by the results obtained. The successful integration of data on a massive scale—more than 976 million entities and 1.1 billion relationships—and the system's ability to respond to complex investigative queries, such as discovering asset hiding paths with up to eight degrees of separation, serve as direct proof that the ecosystem can overcome data fragmentation and enhance analytical capabilities. The most profound impact of the ecosystem is the creation of a unified analytical plan for the AGU, which fundamentally changes how investigators interact with data.

In this vein, the ecosystem positions itself as a powerful tool for the AGU and other state bodies (Public Prosecutor's Office, Courts of Auditors) to increase the recovery of public funds, shift from a reactive to a proactive investigative stance, and promote a data-driven decision-making culture in public security. For computer science and knowledge engineering, this work provides a large-scale, real-world case study demonstrating the power of semantic technologies to solve complex data integration problems. The proposed methodologies offer a model for researchers and practitioners in similar domains.

9.1 STUDY LIMITATIONS AND FUTURE PERSPECTIVES

The main limitations of this study include: the fact that the models and ontology were developed and validated using Brazilian government datasets, and their direct applicability to other legal and administrative systems would require adaptation; Issues related to the quality of source data remain, since the effectiveness of the ecosystem is inherently dependent on the quality of the input data, and inconsistencies and errors in the source databases remain a challenge; and the deliberate choice of explainable models has led to an underutilization of more complex and potentially more powerful, but less transparent, techniques.

Based on the identified limitations and the need for continuous evolution of the proposed ecosystem to address the dynamic nature of financial fraud, the following research lines are envisioned for future work:

- Validation with Unstructured Data and OSINT (Complementary studies currently under development): Currently, master's theses are underway to validate and expand the ecosystem's capacity to handle unstructured data and open-source intelligence (OSINT). These studies seek to automatically integrate information from social networks with structured

government records, overcoming the data fusion gap identified in this thesis.

- LLM to discover entities: Investigating the use of Large Language Models (LLMs) tuned for the semi-automated extraction of entities and relationships from unstructured data sources, such as legal documents and court records, to enrich the Knowledge Graph further.

- "Ecosystem-as-a-Service"(EaaS) Platform: Aiming for technological sovereignty and scalability, a future work front is the development of a reusable platform that can be easily instantiated by other control and oversight bodies (such as the Public Prosecutor's Office and the Courts of Auditors). The focus would be on reducing barriers to entry for the use of graph technologies in the public sector and automating the governance and data lineage steps that were methodologically grounded in this research.

- Implementation of GNNs for Proactive Anomaly Detection: While this thesis established a solid foundation through coded heuristics and explainable inference rules, a natural next step is to migrate to Graph Neural Network (GNN) architectures, such as Heterogeneous Graph Transformers (HGT) or Graph Attention Networks (GAT). With increased data maturity and the creation of a repository of labels from real cases confirmed by the AGU, it will be possible to train models capable of identifying non-trivial fraud patterns that overcome the rigidity of expert knowledge-based heuristics.

- Improvement of Explainable AI (XAI) for Evidentiary Purposes: Given that the ecosystem's results aim to support judicial asset-seizure measures, the development of specialized XAI techniques for graphs is fundamental. Future work could focus on creating mechanisms that translate neural network weights or graph connections into narrative provenance reports, ensuring that the generated intelligence is not only actionable but entirely defensible and auditable in court.

REFERENCES

ABOWD, G. D.; DEY, A. K.; BROWN, P. J.; DAVIES, N.; SMITH, M.; STEGGLES, P. Towards a better understanding of context and context-awareness. In: SPRINGER. *International symposium on handheld and ubiquitous computing*. [S.l.], 1999. p. 304–307. Cited in page 19.

ABROUK, L.; CHERGUI, H.; AHAGGACH, H. Ontofic: an ontology for financial fraud detection and customer behavior modeling. In: *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*. [S.l.: s.n.], 2023. p. 509–513. Cited 6 times in pages ix, 14, 15, 16, 103, and 104.

AKBIK, A.; BLYTHE, D.; VOLLGRAF, R. Contextual string embeddings for sequence labeling. In: *Proceedings of the 27th international conference on computational linguistics*. [S.l.: s.n.], 2018. p. 1638–1649. Cited in page 37.

ALBUQUERQUE, E. O.; ALMEIDA, W. G. de; PRACIANO, B. J. G.; MEDEIROS, M. B. de; MENDONÇA, F. L. L. de; ALBUQUERQUE, R. de O. Data pipelines implementation and management for data engineering: A case study applied to the public sector. In: SPRINGER. *International Conference on Information Technology & Systems*. [S.l.], 2025. p. 212–222. Cited in page 134.

ALGOABRA, M. *Structured Knowledge Extraction from Text using Large Language Models*. Tese (Doutorado) — Universität Rostock, 2024. Cited in page 27.

APACHE NIFI TEAM. *Apache NiFi Overview*. 2025. Available online: <https://nifi.apache.org/components/> (accessed on 15 July 2024). Cited in page 34.

APACHE SOFTWARE FOUNDATION. *Apache Nifi*. 2025. Available online: <https://nifi.apache.org/> (accessed on 15 July 2024). Disponível em: <https://nifi.apache.org/>. Cited in page 33.

BAE SYSTEMS. *The smart way to combat financial crime*. 2024. Available online: <https://www.baesystems.com/en-media/uploadFile/20231215143816/1434557134903.pdf> (accessed on 15 July 2024). Cited in page 41.

BENNETT, M. The financial industry business ontology: Best practice for big data. *Journal of Banking Regulation*, Springer, v. 14, n. 3, p. 255–268, 2013. Cited 3 times in pages 14, 103, and 104.

BIRD, S.; KLEIN, E.; LOPER, E. *Natural language processing with Python: analyzing text with the natural language toolkit*. [S.l.]: "O'Reilly Media, Inc.", 2009. Cited in page 36.

BLEIHOLDER, J.; NAUMANN, F. Data fusion. *ACM computing surveys (CSUR)*, ACM New York, NY, USA, v. 41, n. 1, p. 1–41, 2009. Cited in page 25.

BORST, W. N. *Construction of engineering ontologies for knowledge sharing and reuse*. Tese (Doutorado) — Institute for Telematica and Information Technology, University of Twente, Enschede, The Netherlands, 1997. Cited in page 13.

BRAZIL. Lei nº 12.651, de 25 de maio de 2012. adotar número único para os documentos que especifica e para estabelecer o cadastro de pessoas físicas (cpf) como número suficiente para identificação do cidadão nos bancos de dados de serviços públicos. *Diário Oficial [da] República Federativa do Brasil*, Brasília, DF, 2023. ISSN 1677-7042. Available online: <https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2023/lei/l14534.htm> (accessed on 15 July 2024). Cited in page 139.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001. Cited in page 143.

BRICKLEY, D.; MILLER, L. *FOAF Vocabulary Specification 0.99*. 2014. Available online: <http://xmlns.com/foaf/spec/20140114.html> (accessed on 31 August 2025). Cited 4 times in pages 16, 18, 103, and 104.

BRUN, J.-P.; SOTIROPOULOU, A.; GRAY, L.; SCOTT, C. *Asset recovery handbook: a guide for practitioners*. [S.l.]: World Bank Publications, 2021. Cited 4 times in pages vii, 1, 10, and 11.

CABOT, P.-L. H.; NAVIGLI, R. Rebel: Relation extraction by end-to-end language generation. In: *Findings of the association for computational linguistics: emnlp 2021*. [S.l.: s.n.], 2021. p. 2370–2381. Cited in page 28.

CAPPELLI, M. A.; SERUGENDO, G. D. M. Methodological exploration of ontology generation with a dedicated large language model. *Electronics*, v. 14, n. 14, 2025. ISSN 2079-9292. Cited in page 21.

CARVALHO, V. A. de; ALMEIDA, J. P. A.; GUIZZARDI, G. Using reference domain ontologies to define the real-world semantics of domain-specific languages. In: SPRINGER. *Advanced Information Systems Engineering: 26th International Conference, CAiSE 2014, Thessaloniki, Greece, June 16-20, 2014. Proceedings 26*. [S.l.], 2014. p. 488–502. Cited 5 times in pages vii, 16, 17, 103, and 104.

CHAUDHRI, V.; BARU, C.; CHITTAR, N.; DONG, X.; GENESERETH, M.; HENDLER, J.; KALYANPUR, A.; LENAT, D.; SEQUEDA, J.; VRANDEČIĆ, D. *et al.* Knowledge graphs: introduction, history and, perspectives. *AI Magazine*, v. 43, n. 1, p. 17–29, 2022. Cited in page 19.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. [S.l.: s.n.], 2016. p. 785–794. Cited in page 143.

CHRISTEN, P. Geocode matching and privacy preservation. In: SPRINGER. *International Workshop on Privacy, Security, and Trust in KDD*. [S.l.], 2008. p. 7–24. Cited in page 25.

CHRISTEN, P. *The data matching process*. [S.l.]: Springer, 2012. Cited 2 times in pages 25 and 79.

CHUI, C.; GRÜNINGER, M.; WONG, J. An ontology for formal models of kinship. In: IOS PRESS. *Formal Ontology in Information Systems*. [S.l.], 2020. p. 92–106. Cited 4 times in pages 16, 17, 103, and 104.

CORBETT, P.; BOYLE, J. Chemlistem: chemical named entity recognition using recurrent neural networks. *Journal of cheminformatics*, Springer, v. 10, n. 1, p. 59, 2018. Cited in page 24.

CORREIOS. *Tudo sobre CEP*. 2025. Available online: <https://www.correios.com.br/enviar/precisa-de-ajuda/imagens/tudo-sobre-cep> (accessed on 15 Jan 2025). Cited 2 times in pages 117 and 118.

DAMERAU, F. J. A technique for computer detection and correction of spelling errors. *Commun. ACM*, Association for Computing Machinery, New York, NY, USA, v. 7, n. 3, p. 171–176, mar. 1964. ISSN 0001-0782. Disponível em: <https://doi.org/10.1145/363958.363994>. Cited in page 79.

DEPREZ, B.; VANDERSCHUEREN, T.; BAESENS, B.; VERDONCK, T.; VERBEKE, W. Network analytics for anti-money laundering—a systematic literature review and experimental evaluation. *INFORMS Journal on Data Science*, INFORMS, 2025. Cited in page 18.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. Cited in page 24.

EKBAL, A.; BANDYOPADHYAY, S. Named entity recognition using support vector machine: A language independent approach. *International journal of electrical and computer engineering*, v. 4, n. 3, p. 589–604, 2010. Cited in page 24.

ELHASSOUNI, J.; QADI, A. E. Ontology engineering methodologies: State of the art. In: *Proceedings of the 5th International Conference on Big Data and Internet of Things*. Cham: Springer International Publishing, 2022. p. 59–72. Cited in page 20.

ESTER, M.; KRIEGEL, H.-P.; SANDER, J.; XU, X. *et al.* A density-based algorithm for discovering clusters in large spatial databases with noise. In: *kdd*. [S.l.: s.n.], 1996. v. 96, n. 34, p. 226–231. Cited in page 151.

FELLEGI, I. P.; SUNTER, A. B. A theory for record linkage. *Journal of the American Statistical Association*, Taylor & Francis, v. 64, n. 328, p. 1183–1210, 1969. Cited 2 times in pages 25 and 26.

FERNÁNDEZ-LÓPEZ, M.; GÓMEZ-PÉREZ, A. Overview and analysis of methodologies for building ontologies. *The knowledge engineering review*, Cambridge University Press, v. 17, n. 2, p. 129–156, 2002. Cited in page 20.

FERNÁNDEZ-LÓPEZ, M.; GÓMEZ-PÉREZ, A.; JURISTO, N. Methontology: from ontological art towards ontological engineering. In: *Proceedings of the Spring Symposium Series on Ontological Engineering*. [S.l.]: American Asociation for Artificial Intelligence, 1997. Cited in page 21.

GALÁRRAGA, L. A.; TEFLIOUDI, C.; HOSE, K.; SUCHANEK, F. Amie: association rule mining under incomplete evidence in ontological knowledge bases. In: *Proceedings of the 22nd international conference on World Wide Web*. [S.l.: s.n.], 2013. p. 413–422. Cited in page 80.

GENESERETH, M. R.; NILSSON, N. J. *Logical foundations of artificial intelligence*. [S.l.]: Morgan Kaufmann, 1987. Cited in page 13.

GÓMEZ-PÉREZ, A.; ROJAS-AMAYA, M. D. Ontological reengineering for reuse. In: SPRINGER. *International conference on knowledge engineering and knowledge management*. [S.l.], 1999. p. 139–156. Cited in page 21.

GOOGLE MAPS. *Visão geral da API Geocoding*. 2025. Available online: <=>https://developers.google.com/maps/documentation/geocoding/overview?hl=pt-br (accessed on 31 August 2025). Cited in page 145.

GRUBER, T. R. A translation approach to portable ontology specifications. *Knowledge acquisition*, Elsevier, v. 5, n. 2, p. 199–220, 1993. Cited 2 times in pages 13 and 47.

GRüNINGER, M.; FOX, M. Methodology for the design and evaluation of ontologies. 07 1995. Cited in page 20.

GUARINO, N.; OBERLE, D.; STAAB, S. What is an ontology? *Handbook on ontologies*, Springer, p. 1–17, 2009. Cited in page 13.

HAMMERTON, J. Named entity recognition with long short-term memory. In: *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*. [S.l.: s.n.], 2003. p. 172–175. Cited in page 24.

HAMMING, R. W. Error detecting and error correcting codes. *The Bell System Technical Journal*, v. 29, n. 2, p. 147–160, 1950. Cited in page 79.

HERRANZ, J.; NIN, J.; SOLE, M. Optimal symbol alignment distance: A new distance for sequences of symbols. *IEEE Transactions on Knowledge and Data Engineering*, v. 23, n. 10, p. 1541–1554, 2011. Cited in page 79.

HOSMER, D. W.; LEMESHOW, S.; STURDIVANT, R. X. *Applied logistic regression*. [S.l.]: Wiley New York, 2000. Cited in page 142.

HU, Z.; DONG, Y.; WANG, K.; SUN, Y. Heterogeneous graph transformer. In: *Proceedings of the web conference 2020*. [S.l.: s.n.], 2020. p. 2704–2710. Cited 3 times in pages 28, 30, and 31.

INTER-AMERICAN DEVELOPMENT BANK. *Olho com a máquina de lavar*. 2024. Available online: <https://www.iadb.org/pt-br/noticias/olho-com-maquina-de-lavar> (accessed on 15 Jan 2025). Cited in page 1.

IQBAL, R.; MURAD, M. A. A.; MUSTAPHA, A.; SHAREF, N. M. *et al.* An analysis of ontology engineering methodologies: A literature review. *Research journal of applied sciences, engineering and technology*, v. 6, n. 16, p. 2993–3000, 2013. Cited in page 20.

JARO, M. A. Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, ASA Website, v. 84, n. 406, p. 414–420, 1989. Cited in page 79.

JIA, L.; JIN, Y.; LIU, Y.; LV, J. Ontological method for the modeling and management of building component construction process information. *Buildings*, v. 13, n. 8, 2023. ISSN 2075-5309. Cited 3 times in pages 20, 21, and 57.

JIANG, N.; DUAN, F.; CHEN, H.; HUANG, W.; LIU, X. Mafi: Gnn-based multiple aggregators and feature interactions network for fraud detection over heterogeneous graph. *IEEE Transactions on Big Data*, IEEE, v. 8, n. 4, p. 905–919, 2021. Cited in page 31.

JURAFSKY, D.; MARTIN, J. H. *Speech and Language Processing (2nd Edition)*. USA: Prentice-Hall, Inc., 2009. ISBN 0131873210. Cited in page 24.

KEJRIWAL, M. *Domain-specific knowledge graph construction*. [S.l.]: Springer, 2019. Cited 2 times in pages 22 and 79.

KEJRIWAL, M. Knowledge graphs: A practical review of the research landscape. *Information*, MDPI, v. 13, n. 4, p. 161, 2022. Cited in page 19.

KENNEDY, A. Winning the information wars: collecting, sharing and analysing information in asset recovery investigations. *Journal of Financial Crime*, Emerald Group Publishing Limited, v. 14, n. 4, p. 372–404, 2007. Cited in page 2.

KUL, G.; UPADHYAYA, S. A preliminary cyber ontology for insider threats in the financial sector. In: *Proceedings of the 7th ACM CCS International Workshop on Managing Insider Security Threats*. [S.l.: s.n.], 2015. p. 75–78. Cited in page 18.

LENAT, D. B.; GUHA, R. V. *Building Large Knowledge-Based Systems; Representation and Inference in the Cyc Project*. 1st. ed. USA: Addison-Wesley Longman Publishing Co., Inc., 1989. ISBN 0201517523. Cited in page 20.

LEVENSHTEIN, V. Binary codes capable of correcting deletions, insertions, and reversals. *Proceedings of the Soviet physics doklady*, 1966. Cited in page 79.

LF PROJECTS. *Real-Time Intelligence on the Lakehouse Sub-Second Analytics for End Users and Agents at Scale*. 2025. Available online: <https://www.starrocks.io> (accessed on 15 August 2025). Cited in page 134.

LI, Z.; WANG, H.; ZHANG, P.; HUI, P.; HUANG, J.; LIAO, J.; ZHANG, J.; BU, J. Live-streaming fraud detection: A heterogeneous graph neural network approach. In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. [S.l.: s.n.], 2021. p. 3670–3678. Cited 2 times in pages 31 and 32.

LINACRE, R.; LINDSAY, S.; MANASSIS, T.; SLADE, Z.; HEPWORTH, T.; KENNEDY, R.; BOND, A. Splink: Free software for probabilistic record linkage at scale. *International Journal of Population Data Science*, v. 7, n. 3, Aug. 2022. Disponível em: <https://ijpds.org/article/view/1794>. Cited in page 37.

LIU, C.; HE, X.; SU, S.; KANG, M. A graph-based method for chinese address matching. *Transactions in GIS*, Wiley Online Library, v. 27, n. 3, p. 859–876, 2023. Cited in page 145.

LIU, F.-F.; WANG, Z.-J. The study of graininess for tibetan named entity recognition. In: EDP SCIENCES. *ITM Web of Conferences*. [S.l.], 2017. v. 12, p. 01008. Cited in page 23.

LIU, Z.; YANG, M.; WANG, X.; CHEN, Q.; TANG, B.; WANG, Z.; XU, H. Entity recognition from clinical texts via recurrent neural network. *BMC medical informatics and decision making*, Springer, v. 17, p. 53–61, 2017. Cited in page 24.

LÓPEZ, M. F.; PÉREZ, A. G.; AMAYA, M. D. R. Ontology's crossed life cycles. In: SPRINGER. *International Conference on Knowledge Engineering and Knowledge Management*. [S.l.], 2000. p. 65–79. Cited in page 21.

LYU, C.; CHEN, B.; REN, Y.; JI, D. Long short-term memory rnn for biomedical named entity recognition. *BMC bioinformatics*, Springer, v. 18, p. 1–11, 2017. Cited in page 24.

MALTEGO TECHNOLOGIES. *Maltego Graph*. 2025. Available online: <https://www.maltego.com/graph/> (accessed on 15 Feb 2025). Cited in page 39.

MAPBOX. *Geocoding API*. 2025. Available online: <https://docs.mapbox.com/api/search/geocoding/> (accessed on 15 Feb 2025). Cited in page 145.

MINISTRY OF JUSTICE OF BRAZIL. *RT - Dados Biográficos do Programa RIC*. [S.l.], 2015. Available online: <https://www.gov.br/mj/pt-br/acesso-a-informacao/governanca/pdfs/biometria-e-controle/20150331-mj-ric-rt-dados-biograficos-do-programa-ric.pdf/> (accessed on 15 Apr 2025). Cited in page 139.

MONFARDINI, G. K. Q.; SALAMON, J. S.; BARCELLOS, M. P. Use of competency questions in ontology engineering: A survey. In: SPRINGER. *International Conference on Conceptual Modeling*. [S.l.], 2023. p. 45–64. Cited in page 164.

MURPHY, K. P. *et al.* Naive bayes classifiers. *University of British Columbia*, v. 18, n. 60, p. 1–8, 2006. Cited in page 143.

N. HARRIS COMPUTER CORPORATION. *i2 Analyst's Notebook Discover and deliver actionable intelligence*. 2025. Available online: <https://i2group.com/solutions/i2-analysts-notebook> (accessed on 19 August 2025). Cited in page 39.

NADEAU, D.; SEKINE, S. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, John Benjamins, v. 30, n. 1, p. 3–26, 2007. Cited in page 24.

NEO4J Inc. *Neo4j - Uncover Hidden Patterns With Graph*. 2025. Available online: <https://neo4j.com> (accessed on 19 August 2025). Cited 3 times in pages 38, 53, and 134.

NEWCOMBE, H. B.; KENNEDY, J. M. Record linkage: making maximum use of the discriminating power of identifying information. *Communications of the ACM*, ACM New York, NY, USA, v. 5, n. 11, p. 563–566, 1962. Cited in page 25.

NILES, I.; PEASE, A. Towards a standard upper ontology. In: *Proceedings of the international conference on Formal Ontology in Information Systems-Volume 2001*. [S.l.: s.n.], 2001. p. 2–9. Cited 4 times in pages 14, 15, 103, and 104.

NOY, N. *Ontology Development 101: A Guide to Creating Your First Ontology*. [S.l.]: Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and ..., 2001. Cited in page 21.

O TEMPO. *Média de moradores por residência cai para menos de 3 no Brasil, diz Censo*. 2025. Available online: <https://www.otempo.com.br/brasil/2024/10/25/media-de-moradores-por-residencia-cai-para-menos-de-3-no-brasil-> (accessed on 15 Jan 2025). Cited in page 147.

OBERLE, D.; ANKOLEKAR, A.; HITZLER, P.; CIMIANO, P.; SINTEK, M.; KIESEL, M.; MOUGOUIE, B.; BAUMANN, S.; VEMBU, S.; ROMANELLI, M. *et al.* Dolce ergo sumo: On foundational and domain models in the smartweb integrated ontology (swinto). *Journal of Web Semantics*, Elsevier, v. 5, n. 3, p. 156–174, 2007. Cited in page 14.

OLARU, A.; FLOREA, A. M.; SEGHROUCHNI, A. E. F. Graphs and patterns for context-awareness. In: SPRINGER. *Ambient Intelligence-Software and Applications: 2nd International Symposium on Ambient Intelligence (ISAmI 2011)*. [S.l.], 2011. p. 165–172. Cited in page 19.

ORACLE. *Financial Crime and Compliance Management*. 2024. Available online: <https://www.oracle.com/financial-services/aml-financial-crime-compliance/> (accessed on 15 July 2024). Cited in page 41.

PENG, C.; XIA, F.; NASERIPARSA, M.; OSBORNE, F. Knowledge graphs: Opportunities and challenges. *Artificial Intelligence Review*, Springer, v. 56, n. 11, p. 13071–13102, 2023. Cited in page 19.

PERLINE, R. Strong, weak and false inverse power laws. *Statistical Science*, JSTOR, p. 68–88, 2005. Cited in page 147.

QI, P.; ZHANG, Y.; ZHANG, Y.; BOLTON, J.; MANNING, C. D. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*, 2020. Cited in page 37.

QLIK. *What is data fabric? why you need it best practices*. 2025. Available online: <https://www.qlik.com/us/data-management/data-fabric> (accessed on 15 July 2024). Cited 2 times in pages 33 and 34.

QUANTEXA. *Unify Data and Drive AI-Enabled Decisions in Anti-Money Laundering*. 2025. Available online: <https://www.quantexa.com/> (accessed on 15 Jan 2025). Cited 3 times in pages vii, 40, and 41.

QUINLAN, J. R. Improved use of continuous attributes in c4. 5. *Journal of artificial intelligence research*, v. 4, p. 77–90, 1996. Cited in page 142.

RAHM, E.; DO, H. H. *et al.* Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, v. 23, n. 4, p. 3–13, 2000. Cited in page 25.

SCHREIBER, G. Knowledge engineering. *Foundations of artificial intelligence*, Elsevier, v. 3, p. 929–946, 2008. Cited 2 times in pages 12 and 13.

SHAO, Z.; WANG, X.; JI, E.; CHEN, S.; WANG, J. Gnn-eadd: Graph neural network-based e-commerce anomaly detection via dual-stage learning. *IEEE Access*, IEEE, 2025. Cited in page 32.

SHARMAN, J. C. *The money laundry: Regulating criminal finance in the global economy*. [S.l.]: Cornell University Press, 2011. Cited in page 138.

SIMSER, J. Money laundering and asset cloaking techniques. *Journal of Money Laundering Control*, Emerald Group Publishing Limited, v. 11, n. 1, p. 15–24, 2008. Cited in page 1.

SINGHAL, A. *Introducing the Knowledge Graph: things, not strings*. 2012. Available online: <https://blog.google/products/search/introducing-knowledge-graph-things-not/> (accessed on 15 Jan 2025). Cited 2 times in pages 13 and 18.

SOCIAL LINKS. *A Full-cycle OSINT Investigation Platform*. 2025. Available online: <https://sociallinks.io/products/sl-crimewall> (accessed on 15 Jan 2025). Cited in page 35.

SOCIAL SEARCHER. *Social Searcher - Free Social Media Search Engine*. 2025. Available online: <https://www.social-searcher.com/> (accessed on 15 Jan 2025). Cited in page 35.

SONG, S.; ZHANG, N.; HUANG, H. Named entity recognition based on conditional random fields. *Cluster Computing*, Springer, v. 22, p. 5195–5206, 2019. Cited in page 24.

SOXOJ. *Welcome to the Maigret docs*. 2025. Available online: <https://maigret.readthedocs.io/en/latest/> (accessed on 19 August 2025). Cited in page 35.

SPACY. *spaCy 101: Everything you need to know*. 2025. Available online: <https://spacy.io/usage/spacy-101> (accessed on 15 Jan 2025). Cited 3 times in pages ix, 36, and 37.

STANFORD UNIERSITY. *Protege SOFTWARE*. 2025. Available online: <https://protege.stanford.edu/software.php> (accessed on 15 July 2024). Cited in page 38.

STEVENS, R.; MATENTZOGLU, N.; SATTLER, U.; STEVENS, M. A family history knowledge base in owl 2. In: CITESEER. *ORE*. [S.l.], 2014. p. 71–76. Cited 4 times in pages 16, 17, 103, and 104.

STUDER, R.; FENSEL, D.; DECKER, S.; BENJAMINS, V. R. Knowledge engineering: survey and future directions. In: SPRINGER. *German Conference on Knowledge-Based Systems*. [S.l.], 1999. p. 1–23. Cited in page 12.

SURE, Y.; STAAB, S.; STUDER, R. On-to-knowledge methodology (otkm). *Handbook on ontologies*, Springer, p. 117–132, 2004. Cited in page 21.

TAMAŠAUSKAITĖ, G.; GROTH, P. Defining a knowledge graph development process through a systematic review. *ACM Transactions on Software Engineering and Methodology*, ACM New York, NY, v. 32, n. 1, p. 1–40, 2023. Cited 6 times in pages vii, 22, 24, 67, 174, and 179.

TANG, B.; CAO, H.; WU, Y.; JIANG, M.; XU, H. Recognizing clinical entities in hospital discharge summaries using structural support vector machines with word representation features. In: SPRINGER. *BMC medical informatics and decision making*. [S.l.], 2013. v. 13, p. 1–10. Cited in page 24.

TOBLER, W. R. A computer movie simulating urban growth in the detroit region. *Economic geography*, Taylor & Francis, v. 46, n. sup1, p. 234–240, 1970. Cited in page 150.

TOLEDO, A. d. A. L. F. *COMO O IBAMA PERDEU MAIS DE 1 BILHÃO DE REAIS*. 2023. Available online: <https://piaui.folha.uol.com.br/como-o-ibama-perdeu-mais-de-1-bilhao-de-reais/> (accessed on 15 Feb 2025). Cited in page 151.

TORRES, J. A. S.; SANTOS, P. H. D.; SILVA, D. A. D.; VEIGA, C. E. L.; MEDEIROS, M. B.; VERQARA, G. F.; MENDONÇA, F. L. L.; JÚNIOR, R. T. D. S. Using spatial data and cluster analysis to automatically detect non-trivial relationships between environmental transgressors. In: IEEE. *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*. [S.l.], 2022. p. 98–104. Cited 4 times in pages viii, 151, 152, and 153.

TORRES, J. A. S.; SILVA, D. A. da; ALBUQUERQUE, R. de O.; NZE, G. D. A.; OROZCO, A. L. S.; VILLALBA, L. J. G. Ontology development for asset concealment investigation: A methodological approach and case study in asset recovery. *Applied Sciences*, MDPI AG, v. 14, n. 21, 2024. Cited 46 times in pages vii, ix, 1, 14, 15, 16, 17, 18, 20, 21, 57, 58, 59, 60, 61, 62, 63, 64, 65, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 107, 109, 110, 111, 145, 170, 174, and 175.

TRAJANOSKA, M.; STOJANOV, R.; TRAJANOV, D. Enhancing knowledge graph construction using large language models. *arXiv preprint arXiv:2305.04676*, 2023. Cited in page 28.

UNIVERSITY OF BRASILIA. *CHAMADA PÚBLICA SIMPLIFICADA Nº AGU-01/2021 SELEÇÃO DE PESQUISADORES CÓDIGO: 53.071.023"*. 2025. Available online: <https://redes.unb.br/wp-content/uploads/2023/03/Chamada-Publica-Simplificada-AGU-No-01-2021.pdf> (accessed on 15 Jan 2025). Cited in page 91.

UNODC. *Towards a Global Architecture for Asset Recovery*. Uncac conference edition. [S.l.]: The World Bank and UNODC, 2023. Presented at the United Nations Convention against Corruption Third Conference of States Parties, Doha, Qatar, November 9-13, 2009. Cited in page 10.

USCHOLD, M.; GRUNINGER, M. Ontologies: Principles, methods and applications. *The knowledge engineering review*, Cambridge University Press, v. 11, n. 2, p. 93–136, 1996. Cited 2 times in pages 20 and 171.

VEERASEKHARREDDY, B.; RAO, K. S.; KOPPULA, N. An attention based bi-lstm densenet model for named entity recognition in english texts. *Wireless Personal Communications*, Springer, v. 130, n. 2, p. 1435–1448, 2023. Cited in page 25.

VRAHATIS, A. G.; LAZAROS, K.; KOTSIANTIS, S. Graph attention networks: a comprehensive review of methods and applications. *Future Internet*, MDPI, v. 16, n. 9, p. 318, 2024. Cited 2 times in pages 28 and 30.

WEISCHEDEL, R.; PRADHAN, S.; RAMSHAW, L.; PALMER, M.; XUE, N.; MARCUS, M.; TAYLOR, A.; GREENBERG, C.; HOVY, E.; BELVIN, R. *et al.* Ontonotes release 4.0. *LDC2011T03, Philadelphia, Penn.: Linguistic Data Consortium*, v. 17, 2011. Cited 2 times in pages 36 and 37.

WINKLER, W. E. Matching and record linkage. *Wiley interdisciplinary reviews: Computational statistics*, Wiley Online Library, v. 6, n. 5, p. 313–325, 2014. Cited in page 26.

WNęK, K.; BORYłO, P. A data processing and distribution system based on apache nifi. *Photonics*, v. 10, n. 2, 2023. ISSN 2304-6732. Disponível em: <https://www.mdpi.com/2304-6732/10/2/210>. Cited in page 34.

WU, B.; CHAO, K.-M.; LI, Y. Heterogeneous graph neural networks for fraud detection and explanation in supply chain finance. *Information Systems*, Elsevier, v. 121, p. 102335, 2024. Cited in page 31.

WU, Z.; PAN, S.; CHEN, F.; LONG, G.; ZHANG, C.; PHILIP, S. Y. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, IEEE, v. 32, n. 1, p. 4–24, 2020. Cited in page 80.

XIE, W.; HE, J.; HUANG, F.; REN, J. Supply chain financial fraud detection based on graph neural network and knowledge graph. *Tehnički vjesnik*, Sveučilište u Slavonskom Brodu, Stojarski fakultet, v. 31, n. 6, p. 2055–2063, 2024. Cited in page 32.

ZHANG, S.; TONG, H.; XU, J.; MACIEJEWSKI, R. Graph convolutional networks: a comprehensive review. *Computational Social Networks*, Springer, v. 6, n. 1, p. 1–23, 2019. Cited 2 times in pages 28 and 29.

ZHANG, T.; SHAN, H.-R.; LITTLE, M. A. Causal graphsage: A robust graph method for classification based on causal sampling. *Pattern recognition*, Elsevier, v. 128, p. 108696, 2022. Cited 2 times in pages 28 and 29.

ZHANG, Z.; BETHARD, S. A survey on geocoding: algorithms and datasets for toponym resolution. *Language Resources and Evaluation*, Springer, v. 59, n. 2, p. 1775–1796, 2025. Cited in page 145.

ZHOU, A.; XU, X.; RAGHUNATHAN, R.; LAL, A.; GUAN, X.; YU, B.; LI, B. Knowgraph: Knowledge-enabled anomaly detection via logical reasoning on graph data. In: *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*. [S.l.: s.n.], 2024. p. 168–182. Cited in page 32.