Formalizing the DATASUS RTS: An Ontological Model for a Resource Description Framework Knowledge Graph

Vitor Pires[1], Dalvan Griebler[1], Felipe Meneguzzi[1,2]

[1]Pontifical Catholic University of Rio Grande do Sul (PUCRS), Porto Alegre, Brazil
[2]University of Aberdeen, Aberdeen, Scotland

{vitorffpires@gmail.com, dalvan.griebler@pucrs.br, felipe.meneguzzi@abdn.ac.uk}

Abstract

The Brazilian DataSUS platform provides vast health databases in relational formats that, while operationally efficient, lack the robust representation needed for advanced scientific data management, restricting interoperability. In this paper, we develop a knowledge engineering pipeline using Scenario 2 of the NeOn methodology to extract, process, and transform knowledge from the DataSUS Health Terminology Repository into a formal knowledge graph that adheres to World Wide Web Consortium standards. We illustrate the potential of this formalization by showing how the graph captures the domain's complex relationships. The resulting graph comprises over 1.4 million triples, with approximately 700,000 associations generated solely through logical inference. Our pipeline provides a foundational resource that enables advanced structural and semantic querying in Portuguese.

1 Introduction

While numerous initiatives provide open health data, the Big Data context of modern healthcare creates a fragmented, heterogeneous ecosystem spanning electronic health records, administrative databases, and clinical repositories. This heterogeneity creates significant syntactic and semantic barriers, preventing the effective aggregation of data required for advanced analytics and data-driven research. To address these challenges, the community has established guiding frameworks emphasizing that data must be findable, machine-readable, and interoperable to support reuse. In this work, we adhere to the FAIR (Findable, Accessible, Interoperable, and Reusable) principle, which serves as a guideline for good data management and stewardship to facilitate knowledge discovery (Wilkinson et al., 2016). These principles advocate for formal representations that allow systems to automatically interpret and integrate data. We implement this by establishing formal semantic relations using the Resource Description Framework, centralizing information, and modeling connections by inferring relations.

Researchers have already identified and classified diverse health terminologies, mapped semantic linkages, and enabled intricate query associations to enhance data set interoperability. This foundational harmonization work is what enables robust data governance and the successful implementation of Common Data Models (CDMs) (Weeks and Pardee, 2019). Initiatives such as the OHDSI consortium (Hripcsak et al., 2015), which uses the OMOP CDM (Overhage et al., 2012; FitzHenry et al., 2015), illustrate the success of this strategy. By mapping diverse data sources to a single, standard representation based on core, well-classified terminologies (e.g., SNOMED-CT, LOINC), these frameworks ensure harmonization, shifting the burden of integration from the researcher to the data custodian and thereby enabling reliable, large-scale research.

In the Brazilian context, the open data ecosystem managed by the Unified Health System (SUS) serves as a prime example of this scenario. Its platform, DataSUS, aggregates vast nationwide administrative databases, including the Hospital Information System (SIH), Mortality Information System (SIM), and Outpatient Information System (SIA). This ecosystem represents a massive repository of health data, predominantly stored in standard tabular formats. Such a design, while crucial for operational efficiency and public accessibility, relies on rigid tables that limit interoperability to simple, direct matches, making it difficult to discover implicit connections or complex associations across different systems. To mitigate this issue, the Health Terminology Repository (RTS) was legally instituted as the national virtual environment for the management and publication of standardized semantic resources and information models. Despite providing logical relational definitions, the

RTS infrastructure lacks the semantic formalization needed to capture knowledge relations among these concepts.

In this work, we formalize a semantic representation aligned with World Wide Web Consortium (W3C) standards (such as RDF, OWL, and SKOS) to enable interoperability, retrieval, and direct application of this data. We adopt scenario 2 of the NeOn methodology (Gomez-Perez and Suárez-Figueroa, 2009) to re-engineer these non-ontological tabular resources systematically into W3C standards (Suárez-Figueroa et al., 2015). We develop a knowledge engineering pipeline that extracts and processes the existing information base (the RTS) to semantically enrich the DataSUS relational structure, transforming the tabulated data into a formal knowledge graph.

This work is organized as follows. Section 2 presents background and related works. Section 3 describes the pipeline to build the RDF ontological model. Section 4 discusses the results and Section 5 makes the conclusions of the work.

2 Background and Related Work

Knowledge graphs have a flexible structure, purpose-built to integrate heterogeneous data sources into a unified, machine-actionable knowledge base (Hogan et al., 2021). This integration process systematically maps rigid, application-specific tabular schemas into a flexible, graph-based data model. This transformation converts structured data into an interoperable, standards-compliant resource, making the implicit semantics of the original database explicit and enabling its discovery (Michel et al., 2014).

The formal modeling of structured vocabularies, such as thesauri and taxonomies, relies on standards from organizations such as the World Wide Web Consortium, which provide a common data model (Miles and Bechhofer, 2009). The Resource Description Framework (RDF) is one of those standards, a flexible framework that structures data as "triples", composed of a subject connected by a predicate to an object, to define the specific semantic classes and properties that model the repository's business rules. The World Wide Web Consortium also recommends the Simple Knowledge Organization System, designed to represent and publish structured controlled vocabularies, thesauri, and classification schemes. It provides a simple yet expressive set of predicates for asserting the semantic relationships common to terminologies, allowing for the assignment of labels and the organization of terms into hierarchies (Miles and Bechhofer, 2009).

The Web Ontology Language (OWL) enables advanced reasoning by defining complex constraints and axioms that enable consistency checks and the inference of implicit relationships within health records. Knowledge Reasoning engines, specifically the OWL 2 RL profile, actively expand this base by deriving new insights from existing OWL structures. GraphDB serves as the triplestore, housing the serialized knowledge graph and executing all reasoning processes. The platform leverages its inference capabilities to enforce structural robustness and logical consistency, and activates the inference ruleset to expand the graph's knowledge.

Related research validates the efficacy of this formal representation approach in complex domains, such as its use in managing the formal hierarchies of SNOMED CT, illustrating its scalability and semantic adequacy (El-Sappagh et al., 2018). Fernández-Breis et al. (2013) identify heterogeneous clinical data sources as a core interoperability challenge, proposing ontologies with automated reasoning as a semantic web solution. Similarly, Chelsom and Dogar (2019) proposed a method to link structured health records based on ISO 13606 with external knowledge sources using OWL and RDF. Their work validates a dynamic coding approach, which enriches static clinical records at runtime by matching clinical context with coded knowledge sources.

To construct knowledge graphs, the methodologies generally adopt either a top-down or a bottom-up approach. Bottom-up methods use a data-first approach, building solutions from examples extracted from entities and relationships present in unstructured datasets such as electronic medical records (Arsenyan et al., 2024). Alternatively, top-down approaches focus on using formal representations to organize the data. The PheKnowLator framework provides an ecosystem of robust, automated environments for constructing ontologically grounded graphs in accordance with international standards, though these environments often require complex mapping layers to handle disparate inputs (Callahan et al., 2024). Other recent top-down strategies utilize advanced models to automate complex medical ontology mapping, bridging the gap between structured data and RDF formats (Mavridis et al., 2025).

We can evaluate ontology quality by adapting software engineering metrics to quantify characteristics like tangledness and redundancy (Duque-Ramos et al., 2014). We can also use strategies to expand on these structural metrics by integrating automated functional testing and treating competency questions as executable unit tests to interrogate the graph directly. Furthermore, as knowledge graphs are increasingly integrated with advanced machine learning architectures, researchers are now leveraging LLMs to assist in ontology evaluation (Lippolis et al., 2025).

The DataSUS Health Terminology Repository, which serves as the baseline for this work, provides its terminologies and business rules in a relational, tabular format serving as the national virtual environment to manage and publish standardized semantic resources, classifications, and nomenclatures for Health Information Systems (SIS). The repository organizes its data into a strict, three-tiered relational hierarchy: terminology groups, terminologies, and terminology sublevels.

The schema categorizes health registries into ten distinct thematic groups, including Health Actions and Services, National Relations, Administrative rules, and Professionals. The diagram in Figure 1 illustrates the existing data structure, highlighting an organization optimized for database operations but that hides the rich semantic connections this work aims to extract and formalize.

Building on these established paradigms and resources, our methodology applies a deterministic, top-down formalization tailored to the Brazilian Unified Health System. Rather than relying on entity extraction, we developed our solution by directly re-engineering curated DataSUS relational databases into W3C standards. To achieve this, we employ the NeOn methodology to formalize a highly structured relational model directly into an RDF graph.

In the context of Brazilian health data, however, the adoption of these semantic technologies remains limited. Previous work that uses DataSUS inputs (da Silva et al., 2022; Barbalho et al., 2022) predominantly focuses on statistical epidemiology, treating resources like the Health Terminology Repository merely as lookup tables for code translation. Figure 1 illustrates the current relational structure of the RTS.While optimized for storage, it hides the rich semantic connections. There has been no systematic effort to uplift the core DataSUS relational models into a formal, reasoning-capable Knowledge Graph. This work bridges this gap by applying a semantic mapping pipeline to the Brazilian national terminology infrastructure.

3 Pipeline

To transform the DataSUS terminologies into a semantic knowledge resource, we adopt Scenario 2 of the NeOn methodology to re-engineer the non-ontological resources of the DataSUS Health Terminology Repository into a formal ontology. As illustrated in Figure 2, we divide our pipeline into three main phases: (1) Data engineering to extract the implicit conceptual model; (2) defining the formal transformation rules; and (3) generating the final RDF knowledge graph and deductive inference. We center the architectural topology of the knowledge graph strictly on the SUS Procedure class. This aligns the semantic model directly with the operational reality of the national healthcare infrastructure.

The pipeline begins by extracting the underlying conceptual model directly from the non-semantic source files originating from the RTS. The source data is in tabular format, accompanied by a file describing the table layout and the relational schema. A custom `LayoutParser` class parses the files to extract the relational model. This structure aligns with the specifications described in the official RTS Wiki.

During preprocessing, the pipeline normalizes raw tabular inputs to generate stable, semantically uniform resource identifiers (URIs). A parsing function cleans the textual data by removing accents, normalizing case, and stripping special characters via regular expressions. This standardization guarantees referential integrity and structural consistency.

Subsequently, we apply transformation procedures to organize the extracted schema into a semantic graph. We orchestrate these mappings using Python configuration dictionaries combined with explicitly declared ontological rules. The system maps standard relational columns to SKOS concepts and OWL classes. We structure these explicit directives to handle hierarchical parent-child linkages, associative properties, and reified n-ary relation events. Table 1 details the specific transformation rules.

We implemented three ontology design patterns during the conceptualization phase. Firstly, we applied the N-ary relation pattern to represent the

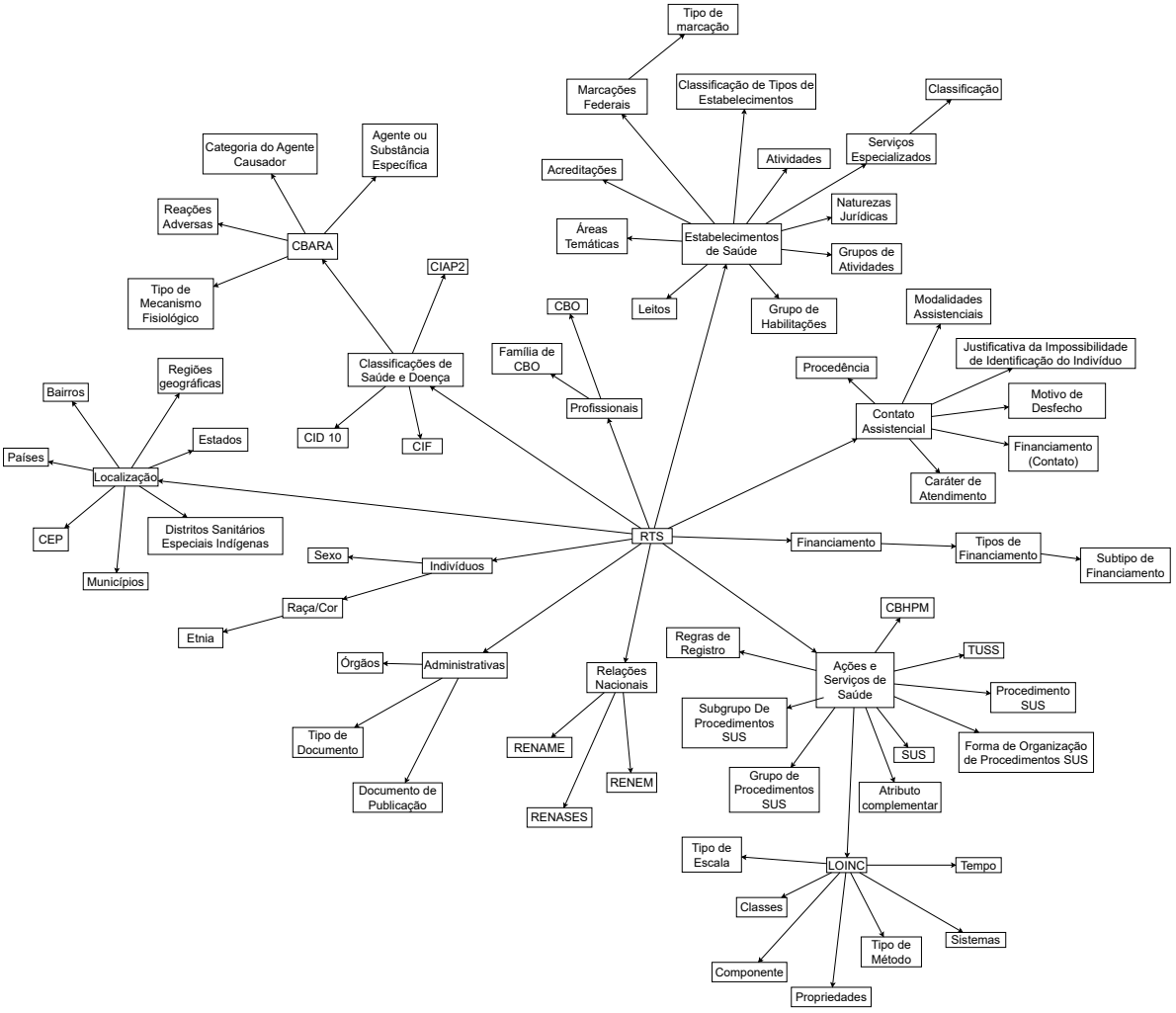

Figure 1: RTS Relational Schema Diagram.

financial increment rules that associate a Procedure with a Federal Marker, while simultaneously assigning specific percentage values. Secondly, we implemented the SKOS concept scheme pattern to instantiate entities as `skos:Concept` within the boundaries of a specific `skos:ConceptScheme`. Furthermore, to satisfy both generalized terminology traversing and domain-specific reasoning, we implemented relationships using standard `skos:broader` properties alongside domain-specific sub-properties. Thirdly, we applied the lexicalization pattern and mapped official system definitions to `skos:prefLabel`, while terminologies, clinical synonyms, and abbreviations were mapped to `skos:altLabel`.

The final phase executes the actual generation of the RDF knowledge graph. The engine applies the predefined transformation rules to construct semantically rich nodes and links, then creates the formal RDF triples and serializes the complete network into the final file. To accomplish this, we used the Resource Description Framework (RDF) to define the specific classes and properties that model the RTS's business rules. We also adopted the Simple Knowledge Organization System (SKOS) as the backbone for managing standard hierarchies, and OWL 2 RL to encode domain logic and expand relations. This process resulted in a formal model comprising 39 main classes, 26 sub-classes, and grouped into 10 categories.

For the graph construction, we consolidated the rules into a central configuration dictionary that dictates how raw tabular columns should be transformed into RDF triples. In these definitions, we established how table rows should be converted into entities, defined the taxonomic relationships to link entities via properties, and instantiated nodes for N-ary relationships to represent their own attributes.

During the transformation phase, specific map-

Transformation Rule	Application & Formalization	Design Pattern
TR1: Entity Formulation	`ont:Procedimento a owl:Class ; rdfs:subClassOf ont:TerminologiaDeProcedimento. ont:TerminologiaDeProcedimento a owl:Class; rdfs:subClassOf skos:Concept.`	Maps core non-ontological tables to unique subjects. All entities are typed as `skos:Concept` and an `owl:Class` subclass.
TR2: Taxonomy Formulation	`ont:eSubgrupoDe a owl:ObjectProperty ; rdfs:subPropertyOf skos:broader ; rdfs:domain ont:Subgrupo ; rdfs:range ont:Grupo.`	Transforms relational hierarchies into semantic taxonomies. Creates rich predicates while maintaining compatibility via `rdfs:subPropertyOf`.
TR3: Associative Formulation	`ont:cidAssociado a owl:ObjectProperty ; rdfs:subPropertyOf skos:related ; rdfs:domain ont:Procedimento ; rdfs:range ont:Cid.`	Creates an association with `skos:related`, utilizing a clear business predicate to represent the domain problem.
TR4: Attribute Formulation	`ont:valorSh a owl:DatatypeProperty ; rdfs:domain ont:Procedimento ; rdfs:range xsd:decimal.`	Transforms descriptive columns into specific datatype properties. Ensures that literals are stored with the precise XSD data type.
TR5: N-ary Event Formulation	`ont:IncrementoFinanceiro a owl:Class. ont:aplicaSeAoProcedimento a owl:ObjectProperty ; rdfs:domain ont:IncrementoFinanceiro ; rdfs:range ont:Procedimento. ont:vlPercentualSh a owl:DatatypeProperty ; rdfs:domain ont:IncrementoFinanceiro ; rdfs:range xsd:decimal.`	Models a table with attributes via reification. The generated Event node links the participating entities and carries the values.

Table 1: Transformation rules for re-engineering the DataSUS Non-Ontological Resources into an RDF Knowledge Graph.

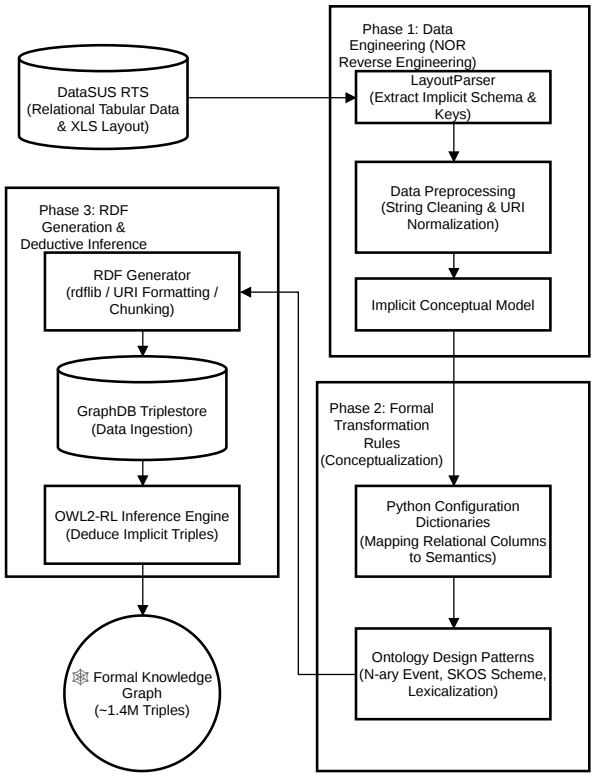

Figure 2: Solution pipeline to build the RDF Ontological Model.

ping challenges were mitigated to preserve graph integrity. Firstly, referential integrity failures occurred when the parser or the historical procedures referenced deprecated tables, professional categories (CBO), or diagnostic codes (CID). To avoid generating orphaned nodes, the URI caching mechanism dynamically validated foreign keys and intentionally dropped unresolvable semantic links. Secondly, we avoided ambiguous literal values in strictly numerical fields, such as financial columns, that contained textual values. The data module avoided conflicts by converting errors to generic `xsd:string` literals, maintaining OWL logical consistency. Finally, unstructured business rules within free-text descriptions, as in `DS_REGRA`, were ingested via the lexicalization pattern as `skos:definition` for future clinical context extraction.

Upon generating the graph structure, we load the serialized triples into the GraphDB. During the ingestion phase, the engine leverages GraphDB's native features to enforce data robustness. The system activates the OWL2-RL inference ruleset to apply the business rules defined in our ontology.

4 Results

The final Knowledge Graph, constructed from the DATASUS RTS terminologies, contains 77 distinct classes, 47 object properties, and 21 datatype properties. The ingestion pipeline populated the graph with 709,144 explicitly stated triples and 715,126 generated triples in the reasoning phase using the OWL2-RL profile, resulting in a final resource comprising 1,424,270 total triples.

Table 2: Quantitative profiling of high-level Group Classes in the RTS Knowledge Graph.

Group Class	Entities	Connections
AcoesServicosSaude	71,670	849,692
ClassificacoesSaudeDoenca	14,394	115,921
Localizacao	6,054	48,432
Profissionais	3,340	34,859
EstabelecimentosSaude	997	11,313
Individuos	273	3,504
RelacoesNacionais	201	1,608
GrupoFinanciamento	58	747
ContatoAssistencial	36	327

Table 3: Distribution of Procedures by High-Level SIGTAP Groups (Hierarchical Aggregation).

SIGTAP Group	Count
Medicamentos (Medicines)	136
Procedimentos cirúrgicos (Surgical Procedures)	94
Procedimentos clínicos (Clinical Procedures)	85
Procedimentos c/ finalidade diagnóstica (Diagnostic)	76
Transplantes de órgãos, tecidos e células (Transplants)	27
Órteses, próteses e materiais especiais (OPM)	24
Ações de promoção e prevenção em saúde (Prevention)	10
Procedimentos para Ofertas de Cuidados Integrados	10
Ações complementares da atenção à saúde	7

We conducted a quantitative profiling of the graph's final state using metadata SPARQL queries. This analysis verified the successful ingestion of the schema and quantified the scale of the populated data. We executed specific `COUNT(DISTINCT ?class)` queries to confirm the `owl:Class` definitions. Furthermore, we executed `GROUP BY ?type` aggregate queries to generate a complete census of instance counts for every domain class. For details on entities and connections in each group class, see Table 2.

Subsequently, we validated the ruleset by executing SPARQL queries targeting the inferred triples. Specifically, a SELECT query for the generic superclass successfully returned instances that the source data explicitly asserted as subclasses. This test showed that the `rdfs:subClassOf` successfully enforced the hierarchy. The original relational structure also has procedures that can be linked to a subgroup or organizational form at varying depths. To validate the graph's ability to handle this structural variance, we executed SPARQL queries using property paths, evaluating the definitions `ont:belongsToGroup` and `ont:belongsToSubgroup`.

Table 3 shows the aggregation of procedures by their high-level Groups. The results confirm that the graph correctly normalizes the hierarchy, allowing for consistent aggregation regardless of the procedure's depth in the tree.

To measure the graph's structural characteristics, we deployed the Ontology Quality Requirements and Evaluation (OQuaRE) framework (Duque-Ramos et al., 2013). Cohesion metrics show that 31 distinct properties converge exclusively on the `ont:Procedimento` class and that it functions as a highly cohesive central group. Regarding taxonomy and vocabulary formulation, we evaluated the depth of the inheritance tree (DITOnto) and the Number of ancestor classes (NACOnto). The DITOnto scored a maximum depth of 2, indicating a relatively flat, broad hierarchical structure. The NACOnto score of 1.09 demonstrates that multiple inheritance is extremely low; the vast majority of concepts inherit from a single, unambiguous parent. This indicates a straightforward, easily navigable taxonomy. Finally, we measured Annotation Richness (ANOnto) at 1.05 annotations per entity. This metric indicates that, on average, each node has slightly more than 1 textual description, highlighting an area for future enrichment.

To compare the semantic model over relational databases, we conducted a comparative query analysis. In the native DataSUS architecture, answering complex intersectional questions requires multi-table `JOIN` operations. For example, retrieving the authorized workforce for a specific diagnosis involves traversing five tables, `TB_CID`, `RL_COMPOSICAO_CID`, `TB_PROCEDIMENTO`, `RL_COMPOSICAO_CBO`, `TB_CBO`. By contrast, the proposed knowledge graph uses property chains and deductive inference. To better illustrate this test, Table 4 evaluates four Competency Questions (CQs), contrasting the relational logic with the semantic approach.

Listing 1 provides the actual SPARQL query used to resolve CQ2, highlighting the efficiency of traversing the structured graph compared to relational SQL.

Listing 1: Example SPARQL query resolving CQ2 for Bariatric CBOs and Funding Values.

```
SELECT DISTINCT ?nomeProc ?valorSh
    ?nomeCbo WHERE {
    ?proc a ont:Procedimento ;
          skos:prefLabel ?nomeProc ;
          ont:valorHospitalar ?valorSh ;
          ont:requerCbo ?cbo .
    ?cbo skos:prefLabel ?nomeCbo .
    FILTER(CONTAINS(LCASE(STR(?nomeProc)),
```

Table 4: Comparative Query Analysis: Relational Database vs. Semantic Knowledge Graph

Competency Question (CQ)	Relational Model Approach	Semantic Model Approach
CQ1 (Clinical Interoperability): Which surgical procedures are associated with Obesity (E66)?	Requires three `JOIN` operations across `TB_CID`, `RL_COMPOSICAO_CID`, and `TB_PROCEDIMENTO`.	Resolved in a single semantic hop using `?proc ont:temCid ?cid`.
CQ2 (Administrative Constraints): Which CBOs are authorized for bariatric surgery, and what is the funding value (VL_SH)?	Requires navigating five distinct tables, including n-ary event tables holding financial increments.	Retrieved by querying properties converging on the central hub: `?proc ont:valorHospitalar ?val ; ont:requerCbo ?cbo.`
CQ3 (Hidden Policy Inference): Which diagnostic codes are implicitly linked to the Oncology policy area?	Requires five `JOIN` operations traversing `TB_CID`, `RL_COMPOSICAO_CID`, `TB_PROCEDIMENTO`, `RL_COMPOSICAO_MARCACAO_FEDERAL`, `TB_MARCACAO_FEDERAL`, and `TB_AREA_TEMATICA`.	Retrieved via OWL2-RL property chains. Query asks for `?cid ont:cidPertenceAreaTematica ?area.`
CQ4 (Lexical Retrieval): What are the terms for clinical classifications related to nutritional monitoring?	Requires `UNION ALL` queries combining `LIKE '%term%'` operators across tables and columns (e.g., `NO_LEIGO_CIAP2`, `DS_SINONIMO_1`, `DS_ABREVIATURA`).	Solved via the `skos:altLabel` property, unifying all lexical variations under a single concept node.

```
        "bariatrica"))
} LIMIT 10
```

To operationalize the knowledge graph for downstream applications, we designed parameterized SPARQL templates that act as dynamic interfaces natively in Portuguese. Leveraging W3C language tags (e.g., `@pt`) and regular expression filters, these templates perform flexible text matching across both official medical classifications (`skos:prefLabel`) and layman synonyms (`skos:altLabel`).

Listing 2 illustrates a core semantic querying template. In a production environment, the placeholder `{{NLP_EXTRACTED_TERM}}` is dynamically replaced by an entity recognizer (e.g., "obesidade"). This automatically expands the search space to include regional acronyms or layman inputs, reducing the syntactic friction between end-users and complex public health databases.

Listing 2: Parameterized SPARQL template for semantic querying in Portuguese.

```
SELECT DISTINCT ?procCodigo ?procNome
    ?cidNome WHERE {
    ?proc a ont:Procedimento ;
          ont:codigo ?procCodigo ;
          skos:prefLabel ?procNome ;
          ont:temCid ?cid .
    ?cid skos:prefLabel|skos:altLabel
        ?cidNome .

    FILTER(REGEX(STR(?cidNome),
        "{{NLP_EXTRACTED_TERM}}", "i")
           &&
               LANGMATCHES(LANG(?cidNome),
               "pt"))
} LIMIT 10
```

5 Conclusion

This work successfully implemented an end-to-end pipeline that transforms the tabular, domain-specific terminologies of the DATASUS RTS into an accessible, logically consistent Knowledge Graph. We began by formalizing the domain's implicit business logic into explicit ontological axioms and structural constraints. By applying an OWL2-RL inference engine to the asserted data, we enriched the graph with over 700,000 new logical triples. Our results validate that the adopted topology, centered on the SUS Procedure, provides consistent information, effectively bridges distinct domains, and enables automated inference of knowledge.

The transition from relational tables to a semantic graph provides a foundational resource for core Natural Language Processing (NLP) research. Entity linking and acronym expansion techniques rely on structured, domain-specific dictionaries to resolve clinical ambiguities; in such cases, our solution can serve as input.

Another example is the development of specialized healthcare question-answering systems with Large Language Models (LLMs), which often hallucinate clinical constraints or funding rules. By utilizing this Knowledge Graph as a factual anchor, an NLP system can extract the normalized entity from the user's prompt, query the graph to retrieve verified, interconnected facts, and inject this context into the LLM.

Despite these advancements, we acknowledge limitations inherent to our approach. First, relying

on a single data source limits the graph's clinical universality. Second, the system's effectiveness is constrained by the quality of the underlying coding maps; inconsistencies in our logic or in the source DATASUS files propagate into the graph unless strictly filtered. Finally, semantic search capabilities require further development.

In future works, we aim to link our graph to international ontologies to enrich the local administrative data. We also propose implementing a Retrieval-Augmented Generation (RAG) architecture, and a promising approach is to extract high-density sub-domains, such as focusing solely on bariatric procedures, for targeted data enrichment. By using the graph's validated structure to ground the outputs of Large Language Models (LLMs), we aim to mitigate hallucinations in health chatbots.

5.1 AI Disclosure

In accordance with the ACL policy on Generative Assistance in Authorship, the authors disclose the use of AI systems to assist in code refactoring, translation of technical terms, and text revision for clarity and grammatical correctness.

5.2 Acknowledgments

This work was partially supported by the Pontifical Catholic University of Rio Grande do Sul (PUCRS) and the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brazil (CAPES) - Financing Code 001. We also had support from FAPERGS 09/2023 PqG (Nº 24/2551-0001400-4), and CNPq Research Program (Nº311012/2025-6).

References

Vahan Arsenyan, Spartak Bughdaryan, Fadi Shaya, Kent Wilson Small, and Davit Shahnazaryan. 2024. Large language models for biomedical knowledge graph construction: Information extraction from EMR notes. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 295–317, Bangkok, Thailand. Association for Computational Linguistics.

Ingridy M. P. Barbalho, Felipe Fernandes, Daniele M. S. Barros, Jailton C. Paiva, Jorge Henriques, Antônio H. F. Morais, Karilany D. Coutinho, Giliate C. Coelho Neto, Arthur Chioro, and Ricardo A. M. Valentim. 2022. Electronic health records in brazil: Prospects and technological challenges. *Frontiers in Public Health*, Volume 10 - 2022.

Tiffany J. Callahan, Ignacio J. Tripodi, Adrianne L. Stefanski, Luca Cappelletti, Sanya B. Taneja, Jordan M. Wyrwa, Elena Casiraghi, Nicolas A. Matentzoglu, Justin Reese, Jonathan C. Silverstein, Charles Tapley Hoyt, Richard D. Boyce, Scott A. Malec, Deepak R. Unni, Marcin P. Joachimiak, Peter N. Robinson, Christopher J. Mungall, Emanuele Cavalleri, Tommaso Fontana, and 13 others. 2024. An open source knowledge graph ecosystem for the life sciences. *Scientific Data*, 11(1):363.

John Chelsom and Naveed Dogar. 2019. Linking Health Records with Knowledge Sources Using OWL and RDF. In *Improving Usability, Safety and Patient Outcomes with Health Information Technology: From Research to Practice*, volume 257 of *Studies in Health Technology and Informatics*, pages 53–58. IOS Press.

Rafaella Alves da Silva, Luiza Gabriela de Araújo Fonseca, João Pedro de Santana Silva, Núbia Maria Freire Vieira Lima, Lucien Peroni Gualdi, and Illia Nadinne Dantas Florentino Lima. 2022. The impact of the strategic action plan to combat chronic non-communicable diseases on hospital admissions and deaths from cardiovascular diseases in Brazil. *PLOS ONE*, 17(6):e0269583.

Astrid Duque-Ramos, Martin Boeker, Ludger Jansen, Stefan Schulz, Miguela Iniesta, and Jesualdo Tomás Fernández-Breis. 2014. Evaluating the good ontology design guideline (goodod) with the ontology quality requirements and evaluation method and metrics (oquare). *PLOS ONE*, 9(8):1–14.

Astrid Duque-Ramos, Jesualdo Fernandez-Breis, Miguela Iniesta-Moreno, Michel Dumontier, Mikel Egaña, Stefan Schulz, Nathalie Aussenac-Gilles, and Robert Stevens. 2013. Evaluation of the oquare framework for ontology quality. *Expert Systems with Applications*, 40:2696–2703.

Shaker El-Sappagh, Francesco Franda, Farman Ali, and Kyung Kwak. 2018. Snomed ct standard ontology based on the ontology for general medical science. *BMC Medical Informatics and Decision Making*, 18.

Jesualdo Tomás Fernández-Breis, José Alberto Maldonado, Mar Marcos, María del Carmen Legaz-García, David Moner, Joaquín Torres-Sospedra, Angel Esteban-Gil, Begoña Martínez-Salvador, and Montserrat Robles. 2013. Leveraging electronic healthcare record standards and semantic web technologies for the identification of patient cohorts. *Journal of the American Medical Informatics Association*, 20(e2):e288–e296.

F FitzHenry, FS Resnic, SL Robbins, J Denton, L Nookala, D Meeker, L Ohno-Machado, and ME Matheny. 2015. Creating a common data model for comparative effectiveness with the observational medical outcomes partnership. *Applied clinical informatics*, 6(3):536—547.

Asuncion Gomez-Perez and Mari Carmen Suárez-Figueroa. 2009. Neon methodology for building ontology networks: a scenario-based methodology.

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia D'Amato, Gerard De Melo, Claudio Gutierrez,

Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. 2021. Knowledge graphs. *ACM Computing Surveys (CSUR)*, 54(4):1–37.

George Hripcsak, Jon D Duke, Nigam H Shah, Christian G Reich, Vojtech Huser, Martijn J Schuemie, Marc A Suchard, Rae Woong Park, Ian Chi Kei Wong, Peter R Rijnbeek, Johan van der Lei, Nicole Pratt, G Niklas Norén, Yu-Chuan Li, Paul E Stang, David Madigan, and Patrick B Ryan. 2015. Observational health data sciences and informatics (ohdsi): Opportunities for observational researchers. *Studies in health technology and informatics*, 216:574–8.

Anna Sofia Lippolis, Mohammad Javad Saeedizade, Robin Keskisärkkä, Aldo Gangemi, Eva Blomqvist, and Andrea Giovanni Nuzzolese. 2025. Large language models assisting ontology evaluation. In *The Semantic Web – ISWC 2025*, volume 16140 of *Lecture Notes in Computer Science*. Springer.

Apostolos Mavridis, Stergios Tegos, Christos Anastasiou, Maria Papoutsoglou, and Georgios Meditskos. 2025. Large language models for intelligent rdf knowledge graph construction: results from medical ontology mapping. *Frontiers in Artificial Intelligence*, Volume 8 - 2025.

Franck Michel, Johan Montagnat, and Catherine Faron Zucker. 2014. A survey of RDB to RDF translation approaches and tools. Research report, I3S. ISRN I3S/RR 2013-04-FR 24 pages.

Alistair Miles and Sean Bechhofer. 2009. Skos simple knowledge organization system reference. W3c recommendation, World Wide Web Consortium (W3C).

J. Marc Overhage, Patrick B. Ryan, Christian G. Reich, Abraham G. Hartzema, and Paul E. Stang. 2012. Validation of a common data model for active safety surveillance research. *Journal of the American Medical Informatics Association*, 19(1):54–60.

Mari Carmen Suárez-Figueroa, Asunción Gómez-Pérez, and Mariano Fernández-López. 2015. The NeOn methodology framework: A scenario-based methodology for ontology development. *Applied Ontology*, 10(2):107–145.

John Weeks and Roy Pardee. 2019. Learning to share health care data: A brief timeline of influential common data models and distributed health data networks in u.s. health care research. *EGEMS (Washington, DC)*, 7(1):4.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, and 34 others. 2016. The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1).