Enabling Clinical Research with Semantic Knowledge Graphs: The Cancer Virtual Lab Platform

Antonella Carbonaro[1,*], Luca Giorgetti[1], Lorenzo Ridolfi[2], Roberto Pasolini[2], Andrea Pagliarani[2], Paolo De Angelis[3], Alice Andalò[3] and Nicola Gentili[3]

[1]*Department of Computer Science and Engineering, Alma Mater Studiorum - University of Bologna (IT)*
[2]*Aivolution Srl, Dogana, San Marino*
[3]*Data Unit, IRCCS Istituto Romagnolo per lo Studio dei Tumori (IRST) "Dino Amadori", 47014 Meldola, Italy*

Abstract

The secondary use of clinical data for research in life sciences is still hindered by data fragmentation, heterogeneity, and limited semantic interoperability across healthcare systems. Semantic technologies and knowledge graphs have emerged as promising enablers to overcome these challenges, yet their adoption in operational research platforms remains limited. In this paper, we present the Cancer Virtual Lab (CVL), a semantic platform designed to enable clinical research through the integration of standardized data representations, biomedical ontologies, and knowledge graph technologies. CVL leverages HL7 FHIR–based data models and RDF/OWL representations to transform heterogeneous real-world oncology data into interoperable, provenance-aware semantic knowledge graphs. The platform has been applied to a large-scale, real-world oncology dataset comprising 36,335 patient records, in which 1,093,705 hospital stay records were successfully converted into 1,151,559 distinct RDF-based FHIR resources. This semantic backbone supports advanced querying, ontology-driven reasoning, and explainable inference over clinical cohorts, enabling reproducible and transparent research workflows. Beyond data integration, CVL provides user-facing tools for researchers and clinicians, including semantic cohort identification, interactive knowledge graph exploration, and natural-language access to clinical data mediated by AI-based agents. Through architectural descriptions and illustrative screenshots, we demonstrate the feasibility and practical impact of semantic knowledge graphs as a foundation for advanced analytics, AI-driven decision support, and large-scale reuse of clinical data in life sciences research.

Keywords

Semantic technologies, Knowledge Graphs, Clinical Research, Oncology, HL7 FHIR, Life Sciences

1. Introduction

Clinical data generated in healthcare practice represent a critical resource for advancing research in life sciences and enabling evidence-based medicine. However, the secondary use of such data remains severely limited by the persistence of clinical data silos, characterized by heterogeneous data models, institution-specific schemas, and fragmented information systems. Even when data are digitally available, they are often locked into proprietary databases or loosely standardized formats, making cross-study analyses, cohort identification, and reproducible research workflows difficult to achieve [1]. Traditional data integration approaches typically rely on ad hoc extract–transform–load pipelines and relational data warehouses, which flatten clinical complexity and provide limited semantic context, hindering meaningful data reuse beyond the original purpose of collection. As a result, without semantic integration mechanisms supporting ontology alignment, reasoning over clinical concepts, and explicit provenance representation, real-world clinical data remain difficult to access, interpret, and reuse at scale, severely constraining transparency, explainability, and trust in clinical research results [2].

Despite the widespread adoption of clinical data standards such as HL7 FHIR and standardized terminologies [3], their impact on clinical research has remained limited. In many real-world settings,

SeWebMeDA-2026: 9th International Workshop on Semantic Web Solutions for Large-scale Biomedical Data Analytics, May 10, 2026, Dubrovnik, Croatia
*Corresponding author.
antonella.carbonaro@unibo.it (A. Carbonaro); a.pagliarani@aivolutionsrl.com (A. Pagliarani); nicola.gentili@irst.emr.it (N. Gentili)

standards are primarily used for data exchange and syntactic interoperability, while research platforms continue to rely on database-centric architectures that offer limited semantic integration and reasoning capabilities. This disconnect can create a gap between standards-compliant data infrastructures and usable research platforms, leaving clinical researchers without accessible tools for semantic querying, cohort identification, and explainable analysis of real-world data [4].

This paper builds upon previously published work describing the semantic architecture and large-scale data integration capabilities of the Cancer Virtual Lab (CVL), and focuses on how these foundations concretely enable clinical research workflows. The main contributions of this paper are as follows. We briefly summarize the CVL semantic integration pipeline, including the transformation of real-world oncology data into RDF-based FHIR resources and ontology-driven knowledge graphs [5], which are treated here as enabling infrastructure providing the semantic foundation for advanced research-oriented functionalities. Building on this foundation, we describe how CVL's semantic knowledge graph can actively support clinical research activities. In particular, we show how the platform enables complex cohort identification, semantic abstraction over heterogeneous clinical variables, and explainable reasoning through ontology-based inference and rule-driven analytics, supporting transparent and reproducible research workflows beyond traditional database-centric approaches. Finally, we present and analyze user-oriented interaction mechanisms that bridge the gap between semantic complexity and research usability, including interactive knowledge graph exploration and natural-language access to clinical data mediated by large language models. This paper emphasizes how these interaction paradigms lower technical barriers, preserve human oversight, and make semantic knowledge graphs practically usable in clinical research contexts. A demonstration of the platform's core functionalities, including semantic cohort identification, knowledge graph exploration, and natural language querying, is available at: https://youtu.be/BHZh5ncL6ZI.

The remainder of this paper is structured as follows. Section 2 provides background on semantic technologies for clinical research and related work. Section 3 outlines the vision and design principles of the CVL platform. Section 4 describes the semantic architecture, including data management, knowledge graph construction, and analytics. Section 5 focuses on user-facing interaction mechanisms, such as interactive exploration and natural-language querying. Finally, Section 6 discusses the results and concludes the paper.

2. Background and Related Work

Ontologies and standardized clinical terminologies play a central role in enabling semantic interoperability in clinical and life sciences data infrastructures [6, 7, 8, 9]. Resources such as SNOMED CT, LOINC, ICD, and the NCI Thesaurus (NCIt) provide controlled vocabularies and formal concept hierarchies that allow clinical entities, observations, diagnoses, procedures, laboratory tests, and biomedical findings to be represented in a consistent and machine-interpretable way [10]. SNOMED CT offers a broad, compositional ontology of clinical concepts and relationships, widely used to encode conditions, procedures, and clinical findings [11]. LOINC focuses on the standardized identification of laboratory and clinical measurements, supporting harmonized representation of test results across institutions [12]. ICD remains the dominant international classification for diagnoses and epidemiological reporting [13], while NCIt provides a domain-focused semantic resource particularly relevant for oncology, including cancers, drugs, biomarkers, and molecular features [14]. When used in combination, these ontologies and terminologies enable semantic normalization and cross-dataset comparability; however, their effective use in research settings requires integration frameworks capable of aligning codes, resolving overlaps, and embedding terminological knowledge into computable semantic models and knowledge graphs [15].

HL7 FHIR (Fast Healthcare Interoperability Resources) has emerged as one of the most widely adopted standards for syntactic and structural interoperability in healthcare data exchange [16]. FHIR defines a modular resource-based model for representing clinical entities such as patients, observations, conditions, procedures, medications, and encounters, combined with RESTful interfaces and

standardized serialization formats. Its flexibility and implementation-oriented design have promoted broad adoption across healthcare systems and research infrastructures. Beyond primary clinical data exchange, FHIR is increasingly used to support secondary use scenarios, including clinical research, real-world evidence generation, and data sharing across institutional boundaries. However, while FHIR provides a common structural model and controlled bindings to terminologies, it does not by itself guarantee deep semantic interoperability or reasoning-ready representations. For advanced research use cases, FHIR resources are often further transformed into RDF and linked with ontologies and knowledge graphs, enabling richer semantic integration, cross-domain querying, and inference [17]. Current research and practice therefore increasingly explore FHIR-based semantic layers that extend interoperability from data exchange to computable knowledge representation and research-oriented analytics.

Knowledge graphs have gained increasing relevance in biomedical and clinical research as a semantic technology for integrating heterogeneous data, domain knowledge, and analytical context into a unified, machine-interpretable framework [18, 19, 20]. By representing entities and their relationships as typed, linked structures grounded in ontologies, knowledge graphs enable semantic integration across diverse sources, including clinical records, biomedical databases, guidelines, and literature-derived knowledge. In the life sciences domain, knowledge graphs have been used to support tasks such as drug repurposing, biomarker discovery, cohort identification, and translational research, by enabling concept-based querying and relationship-driven analytics that go beyond attribute-level data access. Their graph structure naturally supports multi-hop reasoning, semantic similarity, and explainable inference, making them particularly suitable for research scenarios requiring transparency and conceptual abstraction. Recent approaches increasingly combine knowledge graphs with interoperability standards and controlled terminologies, as well as with AI and machine learning methods, to provide hybrid, semantically grounded analytical environments. However, many existing biomedical knowledge graphs remain either knowledge-centric without tight integration with real-world clinical data, or data-centric without sufficient semantic and user-oriented interaction support, highlighting the need for operational semantic platforms tailored to clinical research workflows [21].

In addition to knowledge graph–based approaches, alternative data integration paradigms such as relational data warehouses and OMOP Common Data Model (CDM)–based systems are widely adopted in clinical research [22]. Relational and OMOP-based infrastructures provide standardized schemas, strong support for cohort identification, and mature analytical ecosystems, particularly for observational studies and large-scale real-world evidence generation. However, these approaches primarily operate at the structural and syntactic level, relying on predefined table schemas and code mappings, with limited support for explicit semantic relationships, ontology-driven inference, and provenance-aware reasoning.

Existing efforts to operationalize semantic and knowledge graph–based approaches in biomedical research and clinical data integration span a range of initiatives, from frameworks for disease-specific semantic knowledge graphs to general biomedical knowledge graph ecosystems. For example, disease-centric platforms demonstrate automated semantic integration of structured and literature-extracted biomedical knowledge, enabling query and retrieval of disease-relevant concepts across multiple conditions (e.g., lung cancer, dementia) [23, 24]. More general biomedical KG ecosystems like KG-Hub provide modular tooling for constructing and exchanging biological knowledge graphs compliant with shared models such as the Biolink Model, facilitating reuse and interoperable integration across diverse data sources [25]. Large-scale domain-agnostic biomedical KGs such as Petagraph integrate millions of entities and relations to support cross-domain analytics [26].

Despite these advances, existing platforms exhibit limitations when evaluated against the requirements of clinical research workflows. Many systems focus either on literature or structured reference data rather than tightly integrating real-world clinical records, which are essential for observational studies and cohort discovery. Disease-specific KG frameworks often require bespoke pipelines for each condition and may not generalize across therapeutic areas without significant reengineering [27]. Similarly, while KG-focused ecosystems promote interoperability at the model level, they do not always provide integrated support for ontology alignment, explainable inference over clinical events, or user-

oriented research interfaces that are necessary to support reproducible, secondary research. Moreover, variations in modeling practices and lack of consistent standards across biomedical KGs hinder comparability and reuse of KG assets in translational research. These limitations highlight the need for platforms that combine semantics, standardized clinical data, reasoning, and research-oriented interaction in an integrated manner to support everyday clinical research tasks.

Recent review studies highlight the rapid growth and potential of semantic technologies and knowledge graphs within healthcare and biomedical research. Comprehensive surveys of healthcare knowledge graphs have documented their utility in structuring and integrating heterogeneous biomedical knowledge, supporting applications that span basic science, clinical decision support, pharmaceutical research, and public health, and even exploring synergies with large language models for richer representations and analytics. Systematic reviews on semantic interoperability emphasize that achieving true semantic integration in clinical systems remains a major challenge, with semantic layers and controlled vocabularies being crucial for overcoming data heterogeneity and enabling meaningful cross-system data reuse. These reviews also observe that, despite the promise of semantic layering, existing efforts often focus on either conceptual frameworks or isolated use cases without sufficiently addressing end-to-end research workflows that include reasoning, explainability, and user-oriented interaction [2, 28].

3. The CVL Platform: Vision and Design Principles

The design principles of CVL are motivated by the intrinsic complexity of oncology research and clinical data, which are heterogeneous, semantically rich, high-impact, and subject to strict interpretability and reproducibility requirements. CVL adopts a standards-first approach in which interoperability standards and biomedical ontologies are treated as foundational design elements. HL7 FHIR provides the core data representation model, while shared clinical terminologies and ontologies are used to ground data in explicit domain semantics. In Cancer Virtual Lab, we applied this methodology to a large-scale, real-world oncology dataset comprising 36,335 anonymized patient records, successfully converting 1,093,705 clinical records into 1,151,559 distinct RDF-based FHIR resource types. The process incorporated syntactic and semantic validation, along with expert review, to ensure technical correctness and clinical relevance. The CVL project has been approved by the IRST Scientific Board on 2024-02-13.

Explainability is a primary design requirement of CVL. All data transformations, semantic mappings, and inference processes are explicitly represented and linked to their underlying sources and domain knowledge. Provenance metadata and traceable inference paths allow researchers to inspect how analytical results are produced, supporting transparency, reproducibility, and trust in research outcomes. CVL is designed to support human-in-the-loop interaction throughout the research process, enabling clinicians and researchers to guide data interpretation, validate inferred knowledge, and refine analytical queries without requiring direct interaction with low-level query languages.

To ensure effective usability and alignment with domain expertise, CVL has been developed following a user-centered, co-design approach that actively involves clinicians and biomedical researchers throughout all stages of the system lifecycle, from requirement elicitation to iterative validation. This process, grounded in Design Thinking and Agile methodologies, enables continuous refinement based on real user feedback and ensures consistency with established clinical workflows. The platform abstracts the complexity of underlying semantic technologies through intuitive interaction mechanisms, including natural language interfaces, visual exploration tools, and guided cohort definition functionalities, allowing users to explore semantic data, define cohorts, and inspect knowledge graph structures without requiring in-depth expertise in semantic web technologies. The human-in-the-loop paradigm is therefore conceived to enhance transparency and user control, rather than to impose technical burden. Importantly, the system has been evaluated in real-world clinical settings through pilot studies and usability assessments involving biomedical researchers, demonstrating its practical applicability and accessibility in supporting clinical research tasks.

CVL is designed to serve a heterogeneous group of target users involved in clinical research. Clinicians can explore and interpret semantically enriched clinical data within familiar conceptual frameworks, clinical researchers can perform cohort identification and exploratory analyses, and data scientists can access standardized, machine-readable knowledge graphs for advanced analytics and model development. By addressing the needs of these complementary user groups, CVL fosters interdisciplinary collaboration and supports end-to-end clinical research workflows.

4. Semantic Architecture of CVL

CVL adopts a layered architecture that separates concerns between data ingestion, semantic representation, knowledge management, and user interaction. Standardized data models, biomedical ontologies, and knowledge graph technologies are integrated into a modular and extensible infrastructure, allowing semantic capabilities to be reused across multiple research scenarios. This architectural decomposition supports interoperability, explainability, and scalability, while enabling the platform to evolve independently of specific data sources or user interfaces. The following sections briefly describe the individual layers of the CVL architecture and their roles in supporting semantic, research-oriented clinical workflows.

4.1. Semantic Data Management Layer

The Semantic Data Management Layer forms the foundation of the CVL architecture, transforming heterogeneous clinical data into interoperable, semantically enriched representations for downstream reasoning and analysis, while supporting their secondary use through standardized models that preserve clinical meaning and contextual information.

At the core of this layer lies a semantic ingestion pipeline that maps source clinical data to HL7 FHIR–compliant resources, represented in RDF to enable native integration with semantic web technologies. Data originating from heterogeneous hospital information systems are first normalized into FHIR logical models and subsequently converted into RDF-based FHIR resources, providing a uniform, machine-readable representation aligned with widely adopted interoperability standards. Consistently with recent evidence showing that SNOMED CT, LOINC, ICD-9/10, UMLS, and RxNorm are among the most commonly adopted terminologies in FHIR-based studies [27], our transformation pipeline relies on these widely used and standardized sources to ensure maximum interoperability and alignment with current best practices. In particular, the mapping process is explicitly addressed through a structured transformation pipeline, previously introduced in our earlier work, which formalizes the end-to-end conversion of heterogeneous oncology data into semantically grounded representations. The mapping is performed in two sequential and formally defined transformations. In the first step, raw tabular data are transformed into FHIR-compliant JSON resources through an ETL procedure driven by declarative mapping rules. These rules establish a precise correspondence between source data fields and FHIR resource elements, ensuring structural alignment with the FHIR data model while simultaneously grounding each attribute, where applicable, in standard healthcare ontologies and controlled vocabularies. This step addresses the lack of explicit semantics and relational structure in flat data formats. In the second step, the FHIR JSON resources are systematically translated into RDF representations according to the FHIR RDF specification. This transformation enables the explicit representation of clinical data as a knowledge graph, preserving both syntactic conformance and semantic interoperability. The conversion is implemented through a Python-based toolchain that combines the fhirtordf library, for parsing and transformation into RDF triples, and rdflib, for graph instantiation and serialization into Turtle format. The resulting RDF graphs, together with the domain ontology modeling the clinical context, are deployed into a triple store, thereby enabling integrated querying, schema-level validation, and ontology-driven inferencing within a unified semantic environment. We started from raw clinical data collected from more than 36,000 patients, comprising over 1 million distinct RDF-based FHIR resource types, which were transformed into a knowledge graph containing approximately 174 million RDF triples, supported by a domain

ontology including 86 OWL classes, more than 100 datatype properties, and over 100 object properties. Among the main ontology classes are, for example, cvl:PhysiologicalHistory, cvl:ActiveSubstance, cvl:Diagnosis, cvl:ECOGStatus, cvl:SymptomAssessment, cvl:AdverseEvent, cvl:AdverseEventEffect, cvl:Biomarker, cvl:Patient, cvl:BodyMeasurement, cvl:Department, cvl:Response, cvl:Administration, cvl:Therapy, cvl:TNM, and cvl:TNMClassification. The effectiveness of the pipeline was strongly dependent on close interdisciplinary collaboration. Clinical experts, data engineers, and ontology specialists jointly contributed to the definition of mapping rules, the validation of semantic consistency, and the preservation of clinical relevance in the transformed data. A comprehensive and detailed description of all pipeline stages—including data ingestion, refinement, mapping, validation, and graph deployment—is provided in our previous work, to which the reader is referred for a full formalization and implementation details of the mapping process [5].

The framework adopts a multi-layered validation strategy combining ontology-based reasoning, terminology verification, and constraint-based validation. The core ontology (CVL-O), defined in OWL2, encodes formal semantics through axioms such as rdfs:domain, rdfs:range, owl:disjointWith, and owl:equivalentClass, which are enforced using a Description Logic reasoner (GraphDB and Pellet) to detect inconsistencies such as misclassifications, unsatisfiable classes, and invalid type assignments. For instance, a cvl:MedicationAdministration instance linked to a non-active substance would violate the expected range of cvl:administeredSubstance, generating a type conflict identified prior to graph exposure. In parallel, the integration of standardized biomedical terminologies (e.g., SNOMED CT, LOINC, NCIT, MedDRA, ATC) ensures that all referenced codes are valid and active, preserving terminological consistency and interoperability across heterogeneous clinical data sources. In the implemented validation workflow, these mechanisms were operationalised through a systematic validation pipeline applied to the populated CVL-KG. Ontology-level checks were executed to enforce OWL2 semantic constraints and identify logical inconsistencies, including misclassified entities, unsatisfiable classes, and invalid property assertions. Terminology validation was performed by verifying the correctness and activeness of all external codes against their respective reference vocabularies, ensuring semantic fidelity across integrated clinical datasets. In addition, CVL-specific extensions (cvl:) were reviewed to ensure consistent adherence to OWL modelling principles while maintaining a clear distinction from standardized terminologies. Finally, structural validation procedures were applied to the RDF representation derived from HL7 FHIR data, including automated detection of missing rdf:type declarations, undefined resources, and orphaned or disconnected blank nodes introduced during data transformation and pseudonymization processes. These combined validation steps ensured the overall logical, semantic, and structural integrity of the knowledge graph prior to its use for reasoning and downstream analytical tasks.

Complementing reasoning and terminology checks, SHACL constraints are employed to enforce structural and domain-specific requirements over RDF data. SHACL shapes validate cardinality, datatype correctness, and mandatory relationships among clinical entities. For example, a cvl:AdverseEvent must be associated with exactly one patient, include a severity value, and be linked to a therapy; violations such as missing associations or incomplete attributes are automatically detected during validation. This integrated approach ensures robust data quality control prior to reasoning and graph deployment. The overall validation performance of the loaded repository is summarized in Figure 1, which reports the main validation metrics considered in our assessment framework to ensure semantic accuracy and consistency.

4.2. Knowledge Graph Management Layer

The Knowledge Graph Management Layer builds upon the semantically enriched data produced by the Semantic Data Management Layer and provides the core infrastructure for representing, maintaining, and evolving clinical knowledge within the CVL. This layer is responsible for organizing RDF-based FHIR resources and domain knowledge into a coherent, ontology-driven clinical knowledge graph that can be queried, reasoned upon, and reused across research contexts. The chosen platform for

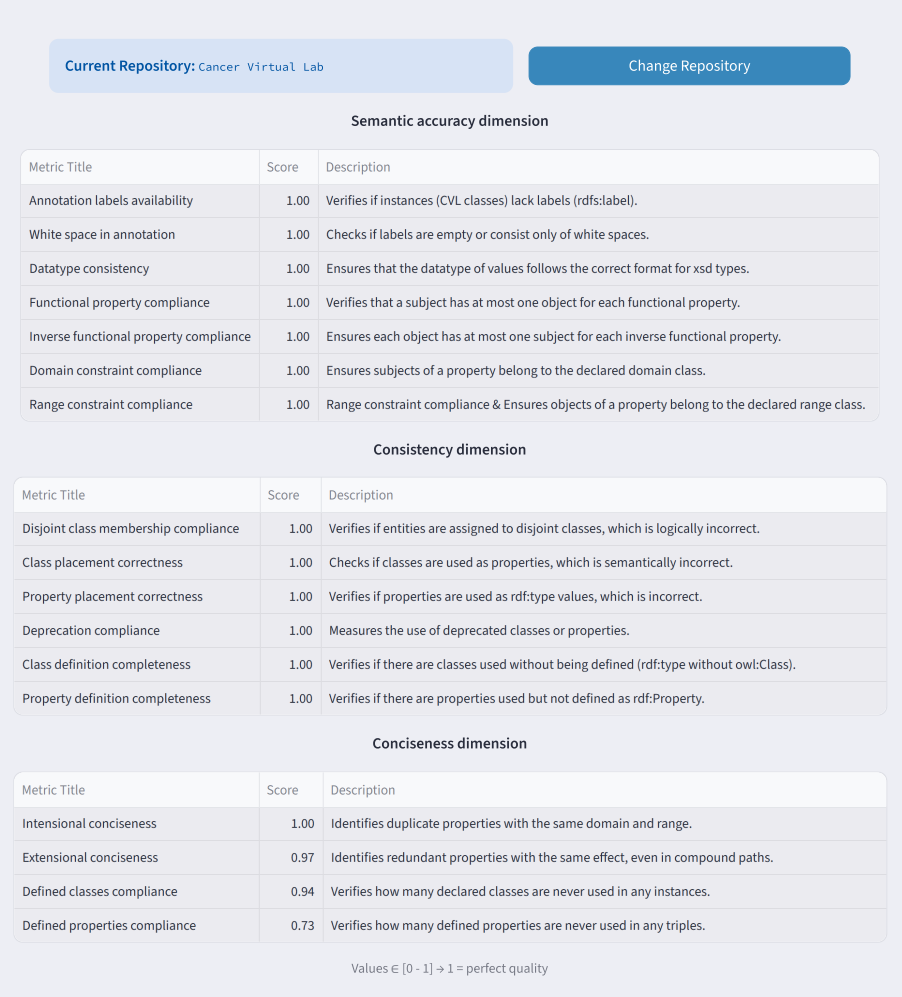

Figure 1: Validation metrics of a loaded repository.

the Cancer Virtual Lab project is Ontotext GraphDB [1]. GraphDB has been selected due to its robust support for standard semantic web technologies, including RDF, RDFS, OWL, and SPARQL,16 ensuring full compatibility with the structured data generated throughout the pipeline.

CVL adopts OWL-based ontological representations to formalize domain semantics and to support logical consistency checking and inference [29]. The knowledge graph integrates instance-level clinical data with schema-level knowledge, allowing clinical observations, procedures, and patient trajectories to be interpreted within a shared semantic framework. In particular, CVL is grounded on the CVL-O ontology, a domain-specific ontology we design to cover the main dimensions of the oncologic care pathway, including patient demographics, diagnoses and staging, therapies and treatment lines, drug administrations and active substances, adverse events and toxicity grading, biomarkers and genomic variants, laboratory tests, clinical responses, body measurements, physiological history, and clinical services. A graphical representation of the ontology structure is provided in Figure 2, illustrating the core classes and relationships that model the comprehensive oncologic care pathway.

By abstracting over underlying storage technologies and exposing the knowledge graph through standardized semantic interfaces, this layer enables uniform access to clinical knowledge for querying and reasoning components. More importantly, it establishes a semantically coherent substrate upon which ontology-driven analytics, explainable inference, and user-facing research tools can operate.

[1] https://graphdb.ontotext.com/documentation/11.0/

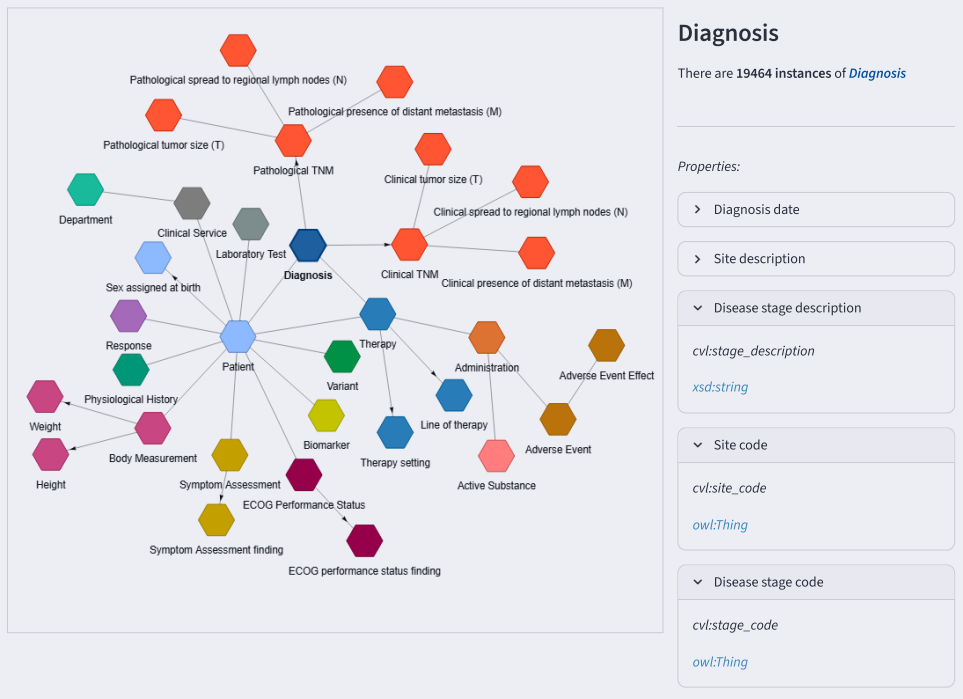

Figure 2: Cancer Virtual Lab Knowledge Graph Ontology.

4.3. Knowledge Graph Analytics and Reasoning

The Knowledge Graph Analytics and Reasoning Layer enables the active exploitation of the CVL knowledge graph for clinical research, leveraging the integrated reasoning engine of GraphDB to infer implicit knowledge from ontology axioms and thereby enrich the underlying dataset. In this work, the built-in rule-based reasoner was employed in combination with a custom rule set to support additional domain-specific inference tasks. In particular, this configuration enabled the derivation of higher-level clinical abstractions, such as cvl:DrinkerPatient and cvl:SmokerPatient, from physiological anamnesis data. GraphDB's reasoning engine is configured through a custom OWL 2 RL profile defined in a proprietary Programmed Inference Engine (.pie) file, which supports a practically useful subset of Description Logic reasoning while maintaining polynomial-time decidability. Within this setting, the reasoner was also used to enforce standard OWL entailments, including owl:equivalentClass, owl:subClassOf, owl:subPropertyOf, owl:inverseOf, as well as rdfs:domain and rdfs:range. Inferred knowledge is explicitly materialized and linked to its underlying evidence, allowing researchers to inspect inference paths, verify assumptions, and assess the provenance of analytical results.

This layer supports explainable clinical research workflows in which analytical outcomes can be traced back to both source data and domain knowledge. This approach enables transparent cohort construction, supports hypothesis generation, and provides a robust foundation for advanced analytics and downstream AI-driven decision support, while preserving interpretability and trust in research results. Figure 3 presents the analytics dashboard for a selected cohort, showcasing how semantic data is dynamically visualized to support survival analysis and other clinical statistics.

A representative reasoning scenario enabled by the CVL architecture can be illustrated through the integration of longitudinal clinical data during neoadjuvant therapy. Patient-level electronic health record data, including laboratory values, treatment dose modifications, and toxicity assessments (e.g., CTCAE grading), are first semantically harmonised through ontology mappings (e.g., LOINC for laboratory tests, CTCAE for adverse events, and SNOMED CT for clinical conditions). Within the CVL knowledge graph, temporal relationships between events are explicitly represented, allowing the system to associate early treatment dynamics with downstream clinical states. For instance, repeated occurrences of high-grade hematologic toxicity, combined with clinically significant dose reductions

Figure 3: Analytics interface of a selected cohort.

and early treatment interruptions within initial therapy cycles, can be semantically linked through predefined clinical rules and temporal constraints to an emerging state of increased risk for suboptimal treatment exposure. This inferred intermediate state is not directly produced by the predictive model itself, but is derived through ontology-based reasoning over structured clinical relationships. The resulting semantic inference is then used to generate an interpretable explanation of the model output, connecting predicted risk trajectories to explicit clinical concepts and temporally grounded evidence encoded in the knowledge graph.

A representative use case of explainable inference in CVL can be illustrated within the toxicity risk prediction task during neoadjuvant therapy. At the knowledge graph level, these features are mapped to semantically grounded clinical concepts (e.g., CTCAE toxicity classes, LOINC-coded laboratory measurements) and contextualised within the patient's longitudinal trajectory. The system then reconstructs a graph-based evidence trail linking the prediction to underlying biomedical evidence, such as temporally correlated toxicity episodes, dose reductions, and treatment interruptions. This evidence trail is exposed to the user as a structured explanation, enabling traceability from the predicted risk score to both data-driven contributions and ontology-aligned clinical concepts.

5. Interaction Mechanism for Clinical Research

The practical impact of semantic technologies and knowledge graphs in enabling interoperable and explainable clinical research ultimately depends on how effectively these capabilities are made accessible to end users. CVL explicitly addresses this challenge by providing user-facing interaction mechanisms designed to make semantic knowledge graphs accessible and usable for clinicians, clinical researchers, and data scientists, without requiring deep expertise in semantic web technologies.

5.1. Interactive Knowledge Graph Exploration and Querying

CVL offers interactive tools for exploring the clinical knowledge graph at different levels of abstraction. Users can navigate clinical entities, relationships, and patient-level data through visual interfaces that expose the underlying semantic structure while hiding low-level implementation details. This enables researchers to inspect how clinical concepts are represented, understand relationships between entities, and verify the semantic consistency of integrated data, while supporting exploratory interaction with

the knowledge graph to facilitate hypothesis generation, cohort validation, and inspection of inferred knowledge.

User-facing interfaces in CVL enable the definition of clinical cohorts through semantically grounded criteria based on high-level concepts aligned with biomedical ontologies, which are automatically translated into executable semantic queries over the knowledge graph. This approach abstracts over heterogeneous coding systems and data representations, enabling reusable cohort definitions and consistent analytical workflows across datasets and studies. Figure 4 depicts the dual interface for semantic cohort selection and interactive graph exploration, allowing users to filter patients based on ontological criteria and inspect individual clinical trajectories.

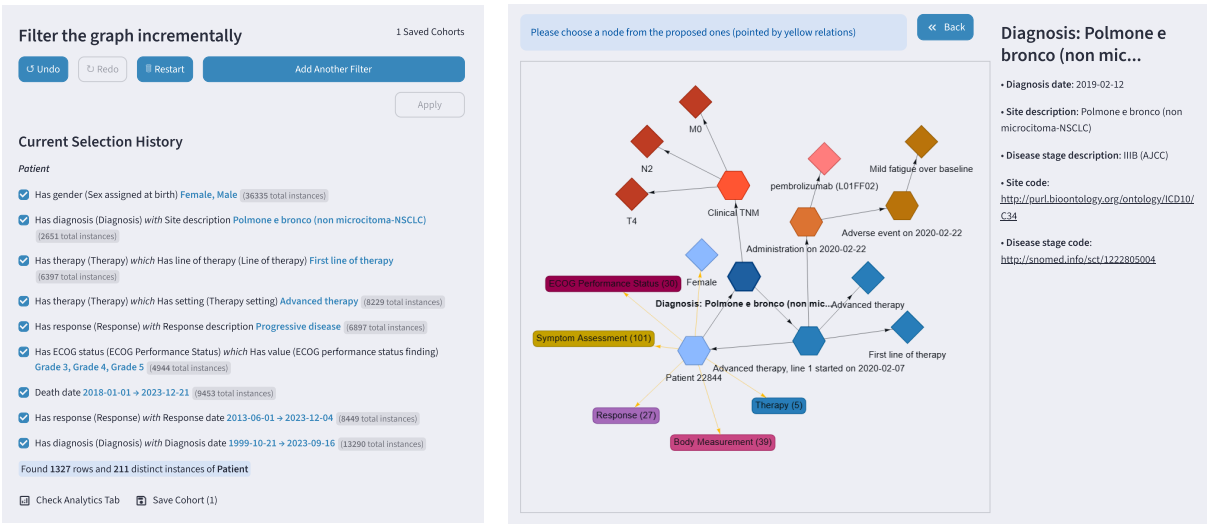

Figure 4: Interactive Knowledge Graph Exploration and Querying.

5.2. Natural Language Access to the Knowledge Graph

To further lower the technical barrier to semantic technologies, CVL integrates natural language interfaces that allow users to query the clinical knowledge graph using domain-specific language. CVL uses a specialized multi-index RAG approach designed specifically for ontology-driven SPARQL generation. The retrieval system is built on ChromaDB and uses Qwen3-Embedding-8B to produce high-dimensional vector representations of clinical terminology and ontological concepts. Instead of storing all contextual knowledge in a single index, which can result in a degradation of the precision of the retrieved context, CVL employs three discrete indices, each compartmentalizing a distinct category of information required by the model to generate a correct SPARQL query. Specifically, three indices are maintained: (i) a curated set of natural language queries paired with validated SPARQL translations, providing syntactic templates for common query patterns; (ii) a collection of textual descriptions of valid multi-hop paths within the graph, capturing semantically coherent traversal patterns (e.g., patient $\longrightarrow$ therapy $\longrightarrow$ active substance $\longrightarrow$ adverse event); and (iii) a repository of ontology-derived class and property definitions, grounding the model in the CVL schema and its associated terminologies. Upon receiving a query, the system performs a K-nearest neighbors retrieval across all indices, selecting a balanced set of examples and contextual fragments to construct a constrained prompt. This prompt enforces strict adherence to the provided schema elements (e.g., class names, property URIs, namespaces) and predefined query structure, effectively limiting the generative model's output space and ensuring the production of syntactically correct and semantically valid SPARQL queries.

Crucially, CVL maintains a human-in-the-loop interaction model: generated queries and results are explicitly exposed to users, who can inspect, refine, and validate them. This design mitigates the risks associated with opaque automation and ensures that natural language access complements expert-driven research workflows. The natural language interface is shown in Figure 5, showing how

a clinical question is converted into a query plan and executed to retrieve specific patient counts and trends.

Figure 5: Natural Language Query Interface.

Across all user-facing components, CVL emphasizes explainability and reproducibility. Query results, inferred knowledge, and analytical outputs are linked to their underlying data sources, ontological definitions, and reasoning rules. This enables users to trace results back to both clinical evidence and domain knowledge, supporting transparent interpretation and reuse of research findings.

6. Discussion and Conclusion

In this paper, we presented the Cancer Virtual Lab as a semantic research platform that operationalizes knowledge graph technologies to support clinical research in oncology. Building on a standards-based semantic integration infrastructure, the focus of this work was on demonstrating how ontology-driven querying, explainable reasoning, and user-facing interaction mechanisms enable transparent and reproducible research workflows over real-world clinical data. Through architectural descriptions, an end-to-end research example, and deployment on large-scale oncology data, we showed how semantic knowledge graphs can bridge the gap between standards adoption and usable research platforms. The experience gained from deploying CVL in a real-world oncology setting confirms that semantic technologies can effectively support complex research tasks, such as cohort identification and exploratory analysis, while preserving explainability, provenance, and human oversight. At the same time, it highlights the importance of research-oriented usability and interaction paradigms in translating semantic capabilities into practical value for clinicians and researchers. Compared with non-semantic approaches, such as relational data warehouses and data lakes, CVL offers a fundamentally different analytical paradigm. Traditional solutions typically rely on schema-specific queries and ad hoc transformations, which limit portability and obscure clinical meaning. In contrast, semantic integration enables queries and analyses to be expressed in terms of clinical concepts rather than database structures, supporting reusable research logic and consistent interpretation across datasets. Moreover, explicit semantics allow inference and validation mechanisms that are not readily supported in database-centric systems, providing richer analytical capabilities at the cost of increased modeling complexity.

Despite these advantages, the current CVL prototype also presents limitations. Semantic data integration and ontology alignment require substantial upfront effort and domain expertise, and reasoning over large-scale knowledge graphs may introduce performance constraints in certain analytical scenarios. In addition, while user-facing tools lower the barrier to interaction, effective use of semantic technologies still requires conceptual understanding of underlying clinical models and assumptions.

More broadly, the experience gained through the development and deployment of CVL offers several lessons for the oncology research community. First, the adoption of standards alone is insufficient to enable effective clinical research; standards must be embedded within platforms that support semantic integration, reasoning, and user interaction. Second, explainability and provenance are essential design requirements for research-oriented infrastructures, particularly when integrating automated analytics and AI-driven components. Finally, human-in-the-loop interaction remains a critical factor for fostering trust and adoption of semantic technologies in clinical research settings.

Future work will focus on extending CVL toward multi-center and federated research scenarios, enabling semantic interoperability across institutions while preserving data governance constraints. Additional directions include tighter integration with advanced analytics and predictive models, the incorporation of AI-driven decision support within the semantic framework, and further alignment with emerging regulatory initiatives such as the European Health Data Space.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT, Grammarly in order to: Grammar and spelling check, Paraphrase and reword. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] M. Lehne, J. Sass, A. Essenwanger, J. Schepers, S. Thun, Why digital medicine depends on interoperability, NPJ digital medicine 2 (2019) 79.

[2] B. Abu-Salih, M. Al-Qurishi, M. Alweshah, M. Al-Smadi, R. Alfayez, H. Saadeh, Healthcare knowl-

edge graph construction: A systematic review of the state-of-the-art, open issues, and opportunities, Journal of Big Data 10 (2023) 81.
[3] S. N. Duda, N. Kennedy, D. Conway, A. C. Cheng, V. Nguyen, T. Zayas-Cabán, P. A. Harris, Hl7 fhir-based tools and initiatives to support clinical research: a scoping review, Journal of the American Medical Informatics Association 29 (2022) 1642–1653.
[4] H. Dong, M. Falis, W. Whiteley, B. Alex, J. Matterson, S. Ji, et al., Automated clinical coding: what, why, and where we are?, NPJ digital medicine 5 (2022) 159.
[5] A. Carbonaro, L. Giorgetti, L. Ridolfi, R. Pasolini, A. Pagliarani, M. Cavallucci, et al., From raw data to research-ready: A fhir-based transformation pipeline in a real-world oncology setting, Computers in Biology and Medicine 197 (2025) 111051.
[6] B. H. De Mello, S. J. Rigo, C. A. Da Costa, R. da Rosa Righi, B. Donida, M. R. Bez, L. C. Schunke, Semantic interoperability in health records standards: a systematic literature review, Health and technology 12 (2022) 255–272.
[7] A. Marfoglia, F. Nardini, V. A. Arcobelli, S. Moscato, S. Mellone, A. Carbonaro, Towards real-world clinical data standardization: A modular fhir-driven transformation pipeline to enhance semantic interoperability in healthcare, Computers in Biology and Medicine 187 (2025) 109745.
[8] A. Luschi, C. Petraccone, G. Fico, L. Pecchia, E. Iadanza, Semantic ontologies for complex healthcare structures: a scoping review, IEEE Access 11 (2023) 19228–19246.
[9] A. Gyrard, S. Abedian, P. Gribbon, G. Manias, R. van Nuland, K. Zatloukal, et al., Lessons learned from european health data projects with cancer use cases: implementation of health standards and internet of things semantic interoperability, Journal of Medical Internet Research 27 (2025) e66273.
[10] W. Hersh, Healthcare data and exchange standards, in: Clinical Informatics Study Guide: Text and Review, Springer International Publishing, Cham, 2022, pp. 185–202.
[11] P. L. Elkin, S. H. Brown, Snomed ct, in: Terminology, Ontology and their Implementations, Springer International Publishing, Cham, 2023, pp. 269–279.
[12] R. L. Richesson, T. R. Campion, B. M. Knosp, D. A. Hanauer, Loinc implementation approaches in academic medical research centers–results from a survey of ctsa sites, Journal of clinical and translational science 9 (2025).
[13] R. Kaur, J. A. Ginige, O. Obst, Ai-based icd coding and classification approaches using discharge summaries: A systematic literature review, Expert Systems with Applications 213 (2023) 118997.
[14] S. De Coronado, L. Remennik, P. L. Elkin, National cancer institute thesaurus (ncit), in: Terminology, Ontology and their Implementations, Springer International Publishing, Cham, 2023, pp. 395–441.
[15] E. Adel, S. El-Sappagh, S. Barakat, K. S. Kwak, M. Elmogy, Semantic architecture for interoperability in distributed healthcare systems, IEEE Access 10 (2022) 126161–126179.
[16] C. N. Vorisek, M. Lehne, S. A. I. Klopfenstein, P. J. Mayer, A. Bartschke, T. Haese, S. Thun, Fast healthcare interoperability resources (fhir) for interoperability in health research: systematic review, JMIR medical informatics 10 (2022) e35724.
[17] A. Carbonaro, A. Marfoglia, F. Nardini, S. Mellone, Connected: leveraging digital twins and personal knowledge graphs in healthcare digitalization, Frontiers in Digital Health 5 (2023) 1322428.
[18] H. Cui, J. Lu, R. Xu, S. Wang, W. Ma, Y. Yu, et al., A review on knowledge graphs for healthcare: Resources, applications, and promises, Journal of biomedical informatics (2025) 104861.
[19] Y. Zhang, X. Sui, F. Pan, K. Yu, K. Li, S. Tian, et al., A comprehensive large-scale biomedical knowledge graph for ai-powered data-driven biomedical research, Nature Machine Intelligence (2025) 1–13.
[20] A. Marfoglia, C. D'Errico, S. Mellone, A. Carbonaro, A knowledge graph-driven framework for deploying ai-powered patient digital twins, Future Generation Computer Systems (2026) 108380.
[21] F. Cremonesi, V. Planat, V. Kalokyri, H. Kondylakis, T. Sanavia, V. M. M. Resinas, et al., The need for multimodal health data modeling: A practical approach for a federated-learning healthcare platform, Journal of biomedical informatics 141 (2023) 104338.
[22] K. K. Gagalova, M. A. Leon Elizalde, E. Portales-Casamar, M. Görges, What you need to know

before implementing a clinical research data warehouse: Comparative review of integrated data repositories in health care institutions, JMIR Formative Research 4 (2020) e17687. doi:10.2196/17687.

[23] S. Jin, H. Liang, W. Zhang, H. Li, Knowledge graph for breast cancer prevention and treatment: literature-based data analysis study, JMIR Medical Informatics 12 (2024) e52210.

[24] R. Reda, F. Piccinini, A. Carbonaro, Semantic modelling of smart healthcare data, in: Proceedings of SAI Intelligent Systems Conference, Springer International Publishing, Cham, 2018, pp. 399–411.

[25] D. R. Unni, S. A. Moxon, M. Bada, M. Brush, R. Bruskiewich, J. H. Caufield, et al., Biolink model: A universal schema for knowledge graphs in clinical, biomedical, and translational science, Clinical and translational science 15 (2022) 1848–1855.

[26] B. J. Stear, T. Mohseni Ahooyi, J. A. Simmons, C. Kollar, L. Hartman, K. Beigel, et al., Petagraph: A large-scale unifying knowledge graph framework for integrating biomolecular and biomedical data, Scientific Data 11 (2024) 1338.

[27] R. Karki, Y. Gadiya, A. Zaliani, B. Pokharel, N. S. Babaiha, M. Ostaszewski, et al., Knowledge graph generator (kgg): A fully automated workflow for creating disease-specific knowledge graphs, Bioinformatics (2025) btaf383.

[28] R. Xu, P. Jiang, L. Luo, C. Xiao, A. Cross, S. Pan, et al., A survey on unifying large language models and knowledge graphs for biomedicine and healthcare, in: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2, 2025, pp. 6195–6205.

[29] R. Reda, A. Carbonaro, V. de Boer, R. Siebes, R. van der Weerdt, B. Nouwt, L. Daniele, Supporting smart home scenarios using owl and swrl rules, Sensors 22 (2022) 4131.