DOCTORAL THESIS

2023

DemKG: A Unified Knowledge Graph Framework for Multimodal Dementia Research Data Integration

Santiago Timón Reina

PROGRAMA DE DOCTORADO EN SISTEMAS INTELIGENTES

SUPERVISORS:

Dr. Mariano Rincón Zamorano

Dr. Rafael Martínez Tomás

A mi familia,

por hacer que esta investigación haya sido posible.

ABSTRACT

Over the past few decades, biomedicine and neuroscience have seen significant growth due to advances in "omics" technologies like genomics, proteomics, and metabolomics, as well as improvements in advanced imaging and deep phenotyping. These advances have generated large, multimodal datasets, prompting the development of platforms to manage biomedical research data. Nonetheless, the integration of these diverse data sources into actionable outputs remains a technological challenge. Bio-ontologies have emerged as critical tools to address this issue, providing a robust framework for data organization, integration, and interpretation. They have also fostered collaboration and knowledge sharing. Concurrently, advancements in graph database technologies have offered more natural representation models and enhanced querying capabilities for interconnected data.

Knowledge Graphs (KGs) constitute a key development in semantic and knowledge modeling, particularly suited for the biomedical domain. Using a graph-based data model, KGs integrate and manage large-scale data from various sources, such as molecular biology, drugs, and disease characterization databases. These graphs link key biomedical entities and their relationships, overcoming data dispersal issues and enabling unified, comprehensive research. In addition, graph-based analytics and machine learning have become essential tools for analyzing complex biological data.

Despite technological progress, the adoption rate among research groups is low, mainly due to the steep learning curve associated with these advanced technologies. Many implementations are either ad-hoc or domain-specific, restricting broader applicability and data sharing. There is a pressing need for more flexible and universally applicable solutions.

This work focuses on the challenges of multimodal data integration in neuroscience and Dementia research. By utilizing semantic models and Knowledge Graphs, we aim to ease the different aspects of interconnecting data entities and artifacts from different research sources. To achieve this, we examine new frameworks and tools across three studies, marking a shift towards data-intensive and integrative applications.

Across this investigation, we present a semantic framework for aligning multidomain biomedical ontologies with research data and show that current graph database management systems can effectively support data-intensive applications in both research and clinical environments. Building on this, we introduce the DemKG framework, a toolkit that facilitates

Knowledge Graph creation and the incorporation of study-specific data, concentrating on Dementia research. Finally, the framework is applied in multiple real-world scenarios to demonstrate its advantages in leveraging interconnected knowledge, ranging from executing expressive graph queries to employing graph embedding techniques, thereby illustrating its relevance for addressing potential questions of interest in Dementia research.

KEYWORDS: Knowledge Graphs, Biomedical ontologies, Semantic Web, Graph databases, Knowledge Representation, Dementia

ACKNOWLEDGMENTS

This doctoral thesis represents the culmination of an extensive journey, supposedly personal, but which is really the product of the collective effort of numerous people to whom I wish to express my most sincere gratitude.

First of all, I would like to thank my directors, Mariano and Rafa, for their dedication and sustained guidance. To Mariano, for his invaluable guidance over so many years and the opportunities he has given me to grow both professionally and personally. To Rafa, for his methodological perspective and his ability to provide encouragement when it has been most needed.

I also want to acknowledge the support received from Norway, which has become a second home for me. I am grateful to Atle Bjørnerud not only for his guidance and hospitality during my stay at the Rikshospitalet Intervention Center in Oslo, but also for the trust placed in me to extend that experience. To Tormod Fladby, for welcoming me to the Department of Neurology at Akershus University Hospital and allowing me to participate in world-class research.

My family deserves special mention. To my parents, whose unconditional support has been instrumental at every stage of this journey. To Maggie and Marty, for their rigorous supervision and companionship in difficult times. And to Alba, whose support at every level has been indispensable to the completion of this work.

On a broader level, I am grateful to society for its contribution in the form of grants and research funding. In particular, to the NILS Science and Sustainability mobility project, coordinated by the Universidad Complutense de Madrid (ABEL-CM-01-2013), which allowed me a research stay in Norway, significant both personally and professionally. Also, to the European Joint Programme - Neurodegenerative Disease Research (JPND), which has funded the studies Dementia Disease Initiation (DDI) and Precision Medicine Interventions in Alzheimer's Disease (PMI), in which I develop my research work and which have been crucial for this work.

To all of them, thank you from the bottom of my heart.

RESUMEN

En las últimas décadas, campos como la biomedicina y la neurociencia han avanzado notablemente, impulsados tanto por tecnologías "ómicas" —genómica, proteómica, metabolómica— como por mejoras en técnicas de imagen y fenotipado de alta resolución. Estos progresos han generado vastos conjuntos de datos multimodales, lo que ha llevado a la necesidad de crear plataformas especializadas para la gestión de datos biomédicos. La unificación de estos conjuntos de datos para obtener resultados prácticos continúa siendo un desafío tecnológico. Las ontologías biomédicas se han establecido como herramientas esenciales para enfrentar este desafío, ofreciendo un marco sólido para la organización, integración e interpretación de datos. Además, han facilitado la colaboración y el intercambio de conocimiento. De forma simultánea, los avances en tecnologías de bases de datos de grafos han proporcionado modelos de representación más intuitivos y capacidades mejoradas para consultar datos interrelacionados.

Los Grafos de Conocimiento (KGs por sus siglas en inglés) representan un avance crucial en el modelado semántico y de conocimiento, con aplicaciones particularmente relevantes en el ámbito biomédico. Mediante un modelo de datos basado en grafos, los KGs facilitan la integración y gestión de extensos conjuntos de datos provenientes de diversas fuentes, tales como biología molecular, farmacología y bases de datos de enfermedades. Estos grafos enlazan entidades biomédicas importantes y sus respectivas relaciones, mitigando así problemas como la fragmentación de datos y posibilitando un enfoque de investigación más cohesivo e integral. Asimismo, al emplear este tipo de estructuras pueden explotarse junto a métodos analíticos y de aprendizaje automático orientados a grafos, que se están consolidando como herramientas fundamentales para el análisis de datos biológicos complejos.

A pesar del avance tecnológico, la adopción de estas tecnologías en grupos de investigación sigue siendo limitada, en gran parte debido a la empinada curva de aprendizaje que conllevan. Esta situación resulta en implementaciones ad-hoc o circunscritas a dominios específicos, lo que restringe una aplicación más generalizada y el intercambio eficaz de datos. En consecuencia, surge una necesidad creciente de soluciones que sean abiertas, flexibles y de aplicación universal.

Este trabajo aborda los desafíos inherentes a la integración de datos multimodales en los ámbitos de la neurociencia y la investigación sobre demencia. Mediante el uso de modelos semánticos y grafos de conocimiento, aspiramos a lograr una representación más intuitiva de los datos biomédicos. A través de tres estudios, evaluamos herramientas y marcos emergentes con el objetivo de catalizar un cambio hacia aplicaciones que sean tanto integrativas como intensivas en el uso de datos.

A lo largo de la investigación, introducimos un marco semántico para alinear ontologías biomédicas multidominio con datos de investigación y demostramos cómo los sistemas de gestión de bases de datos de grafos pueden facilitar aplicaciones intensivas en el uso de datos en contextos tanto clínicos como de investigación. En este contexto, presentamos DemKG, un conjunto de herramientas diseñadas para simplificar la creación de Grafos de Conocimiento y la integración de datos de investigación, con un enfoque particular en la investigación sobre demencia. Finalmente, implementamos estas herramientas en diversos escenarios prácticos para evidenciar su utilidad en la generación de conocimiento interconectado, desde la realización de consultas de datos expresivas basadas en grafos hasta la aplicación de técnicas de embeddings de grafos, subrayando así su relevancia para abordar cuestiones críticas en la investigación de la demencia.

PALABRAS CLAVE: Grafos de Conocimiento, Ontologías biomédicas, Web Semántica, Bases de datos de Grafos, Representación del Conocimiento, Demencia

AGRADECIMIENTOS

Esta tesis doctoral representa el culmen de un extenso recorrido, supuestamente personal, pero que realmente es producto del esfuerzo colectivo de numerosas personas a las que deseo expresar mi más sincero agradecimiento.

En primer lugar, agradecer a mis directores, Mariano y Rafael, por su dedicación y orientación sostenida. A Mariano, por su invaluable guía durante tantos años y las oportunidades que me ha brindado para crecer tanto profesional como personalmente. A Rafa, por su perspectiva metodológica y su capacidad para infundir ánimo cuando más se ha necesitado.

También quiero reconocer el apoyo recibido desde Noruega, que se ha convertido en un segundo hogar para mí. Agradezco a Atle Bjørnerud no solo su orientación y hospitalidad durante mi estancia en el Intervention Center del Rikshospitalet en Oslo, sino también la confianza depositada en mí para prolongar esa experiencia. A Tormod Fladby, por acogerme en el departamento de Neurología del Hospital Universitario de Akershus y permitirme participar en investigaciones de primer nivel.

Mi familia merece una mención especial. A mis padres, cuyo apoyo incondicional ha sido fundamental en cada etapa de este viaje. A Maggie y Marty, por su supervisión rigurosa y su compañía en momentos difíciles. Y a Alba, cuyo apoyo en todos los niveles ha sido indispensable para la culminación de este trabajo.

En un plano más amplio, agradezco a la sociedad por su contribución en forma de becas y financiamiento para la investigación. En particular, al programa NILS Science and Sustainability mobility project, coordinado por la Universidad Complutense de Madrid (ABEL-CM-01-2013), que me permitió una estancia de investigación en Noruega, significativa tanto a nivel personal como profesional. Asimismo, al programa europeo Joint Programme – Neurodegenerative Disease Research (JPND), que ha financiado los estudios Dementia Disease Initiation (DDI) y Precision Medicine Interventions in Alzheimer's Disease (PMI), en los cuales desarrollo mi labor investigadora y que han sido cruciales para este trabajo.

A todos, gracias de corazón.

Contents

Acronyms

API	Application Programming Interface
BFO	Basic Formal Ontology
CKG	the Clinical Knowledge Graph
CLI	Command-Line Interface
COINS	Collaborative Informatics and Neuroimaging Suite
COWAT	Controlled Oral Word Association Test
CRF	Case Report Form
CSV	Comma-separated values
CT scan	Computed Tomography scan
DDI	Dementia Disease Initiation
DL	Description Logics
DOS-DP	Dead Simple OWL Design Patterns
GDBMS	Graph Database Management System
GDS	Geriatric Depression Score
GO	Gene Ontology
GRAPE	Graph Representation leArning, Predictions, and Evaluation
HPO	Human Phenotype Ontology
ISF	Incremental Semantic Framework
KG	Knowledge Graph

KGTK	Knowledge Graph Toolkit
KGX	Knowledge Graph eXchange
KR	Knowledge Representation
LPG	Labeled Property Graphs
MMSE	Mini-Mental State Examination
MRI	Magnetic Resonance Imaging
MRI	functional Magnetic Resonance Imaging
NCATS	National Center for Advancing Translational Sciences
NCBO	National Center for Biomedical Ontology
OBI	Ontology for Biomedical Investigations
OBO	Open Biological and Biomedical Ontologies
ODK	Ontology Development Kit
OGMS	Ontology for General Medical Science
OWL	Web Ontology Language
PET	Positron Emission Tomography
PMI-AD	Precision Medicine Interventions in Alzheimer's Disease
RDBMS	Relational Database Management System
RDF	Resource Description Framework
REST	*Representational State Transfer*
SPOKE	Scalable Precision Medicine Open Knowledge Engine
TSD	Tjeneste for Sensitive Data
TSV	Tab Separated Values
XNAT	eXtensible Neuroimaging Archive Toolkit
XSD	XML Schema Definition
XTENS	eXTENsible platform for biomedical Science

List of Figures

Chapter 1

Introduction

1.1 Research Context

The emergence of omics technologies, such as genomics, transcriptomics, and proteomics, has had a transformative impact on biomedical research [Manzoni et al., 2018, Misra et al., 2019, Glaab et al., 2021, Sun et al., 2011, Lussier and Liu, 2007, Che and Liu, 2017, Che et al., 2015]. These technologies produce a variety of data types, from DNA sequences to gene expression profiles and protein interactions. Alongside, advances in imaging techniques like magnetic resonance imaging (MRI), functional magnetic resonance imaging (fMRI), and Positron Emission Tomography (PET) modalities offer anatomical and functional data, while clinical data gathering provides patient histories, biomarker levels, and treatment outcomes. Lab tests contribute additional layers of data, such as blood chemistry and cellular pathology, and phenotyping adds observable traits and characteristics. In this context, the field of Dementia research stands as a clear example of a complex, multimodal research environment. Dementia, along with other neurodegenerative diseases, presents a multifaceted landscape where a variety of factors contribute to the onset and progression of the condition. The techniques mentioned above have been applied for years to study a range of factors, including genetic predispositions [Del-Aguila et al., 2018, Leonenko et al., 2019, Creese et al., 2021], cardiovascular health [Brain et al., 2023], immune system responses [Peralta Ramos et al., 2022, McManus, 2022], and environmental influences [Killin et al., 2016, Zhao et al., 2021], typically as part of distinct research endeavors. This multimodal environment offers a comprehensive perspective on intricate biological systems, prompting a more integrated approach to disease understanding, as shown in Figure 1.1.

However, this wealth of data introduces its own set of challenges. In many research settings, it is common to navigate through spreadsheets and files from different modalities, where researchers are often left to manage and process data in an ad-hoc manner. This fragmented approach to data management can lead to inefficiencies and errors, hindering

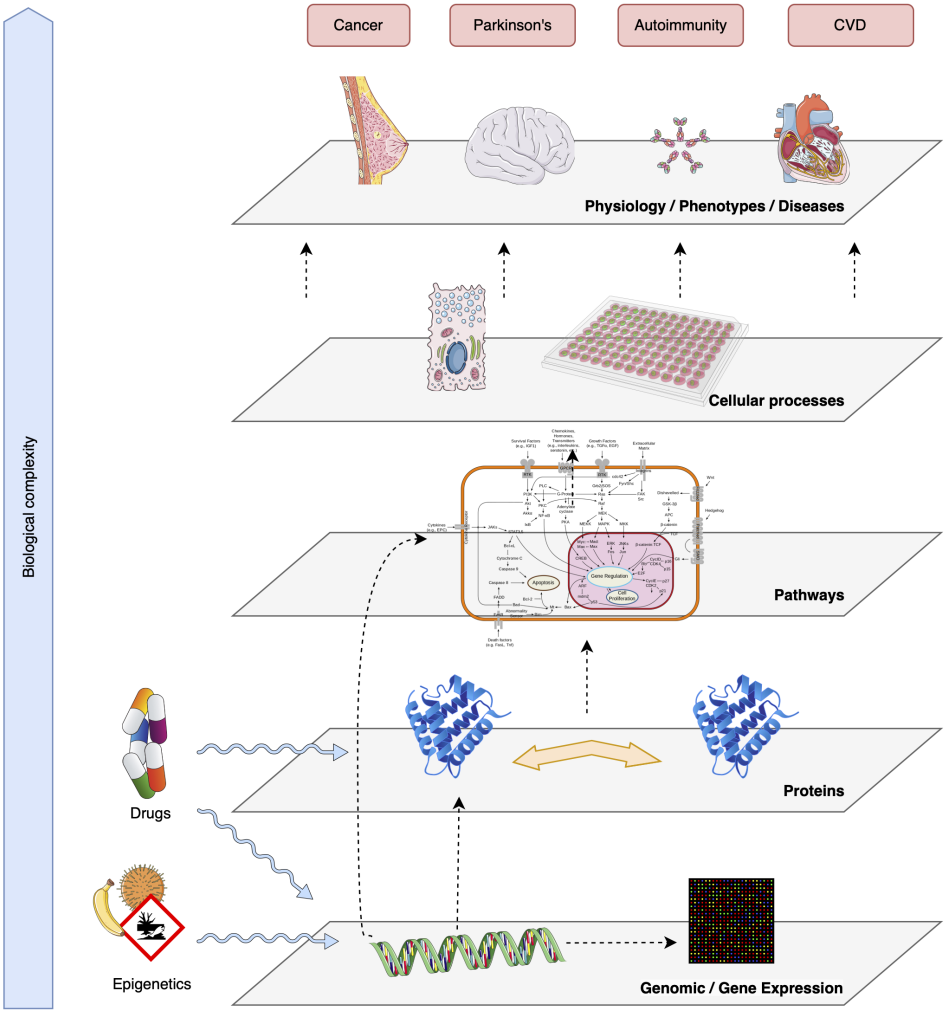


Figure 1.1: A visual representation of the complexity scaling in interactions among biomedical entities.

the potential for integrated analyses. To mitigate these difficulties, several research data management systems have been developed [Izzo, 2016], such as RedCap [Harris et al., 2019] for secure data collection, the eXtensible Neuroimaging Archive Toolkit (XNAT) [Marcus et al., 2007] for neuroimaging data, the Collaborative Informatics and Neuroimaging Suite (COINS) project [Scott et al., 2011] for neuroimaging data storage and retrieval, and the eXTENsible platform for biomedical Science (XTENS) [Corradi et al., 2009, Izzo, 2016] for managing heterogeneous biomedical data. Research data management systems offer the advantage of centralized storage, structured metadata, and streamlined data retrieval, thereby enhancing data integrity, facilitating collaboration, and expediting the analytical process.

Another significant challenge lies in the standardization of data and terminology, which is crucial for interoperability among different data types and research platforms. Inconsistent

nomenclature and varying data formats can impede multimodal data integration, leading to analytical inaccuracies and inefficiencies. Ontologies offer a solution to this problem by clearly defining the shared terminology and relationships between concepts within a specific domain. Their technological evolution is closely related to the evolution and proliferation of the Semantic Web, which developed technologies like Resource Description Framework[1] (RDF) and Web Ontology Language[2] (OWL) to encode these structures, enhancing their accessibility and interpretability.

In biomedicine, bio-ontologies offer standardized vocabularies that improve data representation and structure biomedical knowledge. These ontologies simplify the semantic integration of varied datasets, leading to more accurate analyses. Acting as a shared language, they streamline the merging of data from multiple sources, thereby improving both the quality and utility of the resulting analyses. This role is pivotal for advancing integrative research and contributing to new discoveries in biology and biomedicine. Given their wide adoption in the field, ongoing community efforts aim to maintain and standardize the development of these ontologies. Notable initiatives include the Open Biological and Biomedical Ontologies (OBO) Foundry [Jackson et al., 2021] and the National Center for Biomedical Ontology (NCBO) [Musen et al., 2012], along with its BioPortal [Whetzel et al., 2011].

Ontologies and the Semantic Web are closely related to graph models as a knowledge modeling approach. Formally, a graph $G(V, E)$ is composed of an ordered pair of two disjoint sets: vertices V (also referred to as nodes) and edges E (also referred to as link or relations) [Bollobás, 1998]. This abstraction directly translates concepts and instances into nodes and their relationships into edges. In the context of the semantic web, the RDF graph model employs a subject-object-predicate triple format, naturally resulting in a graph structure where subjects and objects become nodes, and predicates serve as the edges that connect them. Triple-stores, designed to store these RDF triples, act as specialized databases that extend the capabilities of the semantic web for efficient storage and querying. Beyond triple-stores, other graph models like Labeled Property Graphs (LPG) offer additional flexibility, especially for annotating edges. Native graph databases like Neo4j[3] or TigerGraph[4] maintain an explicit graph model, proving efficient querying and data retrieval. This storage approach also enables the application of graph analytics techniques, which can uncover hidden relationships, identify clusters, and perform complex traversals. These techniques enrich the semantic integration and analysis initiated by ontologies and the semantic web,

[1] https://www.w3.org/TR/rdf11-primer/
[2] https://www.w3.org/TR/owl2-primer/
[3] https://neo4j.com/
[4] https://www.tigergraph.com/

serving as a complementary technology that enhances the utility and scalability of integrated analyses across different data modalities.

The rise of large interconnected datasets, coupled with advancements in graph models and databases, has paved the way for the conception of Knowledge Graphs (KGs) [Hogan et al., 2021] and has become another hallmark of semantic and knowledge modeling advances. After Google introduced its Knowledge Graph in 2012[5], highlighting the advantages of the approach, KGs have gained significant traction in industry and academia [Sheth et al., 2019, Ehrlinger and Wöß, 2016]. Conceptually, KGs use a graph-based data model to capture knowledge in application scenarios that involve integrating, managing, and extracting value from diverse sources of data at a large scale [Noy et al., 2019]. Given that biological systems are frequently conceptualized as networks or graphs [Barabasi and Oltvai, 2004, Barabási et al., 2011], KGs are particularly well-suited for this domain. In such graphs, nodes correspond to key biomedical entities, while edges define the various relationships among them. The biomedical field is rich in open databases that offer scientific knowledge from various subdomains, including molecular biology (genomics, proteomics, and pathways), drugs, and disease characterization. These sources hold the potential for a more comprehensive understanding of biomedical phenomena; however, their value is often hindered by their dispersal across different platforms. KGs have emerged as instrumental tools for integrating and exploiting these disparate sources [Nicholson and Greene, 2020], fostering a multitude of projects that aim to unify the spread-out biomedical knowledge. Relevant examples of large biomedical KGs are the Monarch Integrated Knowledge Graph [Mungall et al., 2017], the Clinical Knowledge Graph (CKG) [Santos et al., 2022], PrimeKG [Chandak et al., 2023], and the Scalable Precision Medicine Open Knowledge Engine (SPOKE) [Morris et al., 2023].

With the growing proliferation of KGs, both social and technical challenges have become apparent, notably in the areas of entity naming and graph representation standardization [Badal et al., 2019, Chaves-Fraga et al., 2019]. In response to these challenges, the Biolink Model [Unni et al., 2022] has been developed as a high-level schema offering standardized terminology and relationships for biological entities, facilitating data organization in biomedical KGs. Biolink not only unifies data from diverse sources but also acts as an intermediary between different ontological domains. Parallel to the OBO initiative but focused on KGs, the KG-Hub project [Caufield et al., 2023] contributes a suite of tools and libraries for constructing interoperable KGs, along with mechanisms to encourage their reuse.

In summary, the increasing adoption of KGs, as evidenced by scientific research and various technological initiatives, indicates their expanding role in the biomedical domain.

[5] https://blog.google/products/search/introducing-knowledge-graph-things-not/

This trend presents multiple opportunities for further investigation and contributions to the evolving field.

1.2 Research gaps and hypothesis

As introduced in the previous section, the construction and adoption of KGs in the biomedical domain are rapidly increasing, with several examples of relevance in the context of large, curated biomedical KGs that we further describe below.

- The Monarch Initiative is a collaborative, open science effort that aims to semantically integrate genotype, phenotype, and disease knowledge from a large variety of sources into a KG in support of improved diagnostics and mechanism discovery through various algorithms and tools. Its purpose is to provide a breadth of knowledge unavailable from individual sources and enable diverse user profiles to explore relationships between phenotypes and genotypes across species. Monarch connects phenotypes to genotypes across species by organizing and harmonizing the heterogeneous genotype-phenotype data found across clinical and organism model resources, creating a unified overview of this rich landscape of data sourcesmaking extensive use of bio-ontologies and actively maintaining some, like the Mondo Disease Ontology [Vasilevsky et al., 2022].

- Built on Neo4j, CKG is an open-source platform that integrates a diverse array of omics data, including genomics, transcriptomics, proteomics, and metabolomics, into a single, coherent graph structure. Instead of using a standard data model, CKG opts for an in-house data model, selecting specific concepts and relationships from particular ontologies grouped in self-defined higher-level concepts. Additionally, CKG enhances its KG with integrated statistical and machine learning algorithms to optimize the analysis and interpretation of common proteomics workflows.

- PrimeKG is another multimodal KG aimed at precision medicine analyses. Like its counterparts, it integrates a plethora of resources to describe a broad spectrum of diseases with relationships across major biological scales. PrimeKG sets itself apart from other disease-focused KGs in a few key ways. It offers broader disease coverage, includes detailed drug-disease relationships like indications and contraindications, and spans a wide range of conditions from rare to prevalent, unlike other graphs that may focus exclusively on one or the other. Similarly to CKG, PrimeKG employs a custom data model consisting of 10 node types, some of which include imported ontology terms and 30 types of undirected edges. It is provided in a platform-agnostic CSV format, consisting of triplets that include source nodes, relations, and target nodes.

- SPOKE is a biomedical KG that connects information from 41 biological data sources, structured as 21 different node types and 55 edge types, ranging from low-level molecular biology to pharmacology and clinical practice. It uses 11 different ontologies to organize the data semantically meaningfully and, in its last iteration, also integrates the Biolink model whenever it is found to be practical. Aside from Monarch, SPOKE is one of the most semantically expressive KGs available. It is implemented as a Neo4j database built from a collection of Python scripts and provides a graphical user interface and a REST API for end-user access.

While these examples, along with the literature, show remarkable advantages, we can identify relevant gaps that call for further investigation, especially in the case of their application in Dementia research:

1. **Expressiveness and Standardization**: Most current KGs often resort to less expressive, ad-hoc implementations. These models frequently introduce broader categories defined in-house, without relying on established upper models or reference ontologies. This approach limits the semantic richness and interoperability of the data.

2. **Extensibility and Customization**: Many existing solutions are designed as fixed outcomes, offering limited options for extension or customization. This rigidity impedes the adaptability of these systems to evolving research needs and technological advancements.

3. **Research Data Integration**: There is a noticeable absence of mechanisms for integrating source research data into existing KGs. This shortfall largely stems from the predominant focus of most solutions on serving as external platforms for knowledge retrieval or interaction. However, this limitation is especially relevant for research groups interested in augmenting the knowledge within their datasets or leveraging graph-based analytics and machine learning.

4. **Terminological Limitations in Dementia Research**: Although numerous ontologies aim to encompass Dementia and Neurodegenerative Diseases, recent shifts in the field reveal gaps that require attention to better align with biomarker-based analyses.

5. **Platform Dependency**: Some solutions are tightly coupled with specific platforms, limiting their applicability and hindering their adoption across different computational environments.

To address the aforementioned research gaps, this study formulates the following specific hypotheses:

H1 Employing a more expressive knowledge model that leverages reference ontologies and upper models can significantly improve the semantic richness, interoperability, and utility of biomedical KGs, particularly in the context of Dementia research.

H2 Introducing extensibility and customization features into the design of biomedical KGs will enhance their adaptability, making them more aligned with evolving research and technological landscapes.

H3 The integration of source research data into KGs, facilitated by a robust methodological framework, will lead to more comprehensive and actionable insights in biomedical research.

H4 Addressing the terminological limitations specific to Dementia research through dedicated extensions or adaptations of existing ontologies will result in a more accurate and effective knowledge representation.

H5 The availability of platform-agnostic KG implementations could expand their reach and acceptance for greater integration and collaboration within the biomedical research community.

These hypotheses form the basis for the proposed research, aiming to address the gaps mentioned above through the development and evaluation of a comprehensive, extensible, and semantically rich KG framework tailored for biomedical and, more specifically, Dementia research.

1.3 Research objectives

In accordance with the hypotheses formulated, the main objective of this thesis is to design and implement a modular unifying framework that addresses the identified limitations by providing flexible means for exploiting KGs in the context of Dementia research. This framework will use a solid ontological knowledge model as its foundation by integrating widely accepted biomedical ontologies, employing tools designed to allow flexibility and customizations, and providing means to incorporate research data with a low usage barrier. To approach this endeavor, we define the following research objectives (RO).

RO1 Investigate and evaluate the available biomedical ontologies, considering their scope, implementation, design principles, and ease of alignment in the context of a more extensive knowledge model.

RO2 Identify possible gaps in the target terminology regarding Dementia research.

RO3 Study the general approaches to the problem of harmonizing semantic expressiveness with raw data and propose an integrative framework.

RO4 Determine the most appropriate current technologies for constructing, manipulating, and querying graph models for biomedical research data.

RO5 Design a flexible means to integrate the outputs of the previous objectives to produce biomedical KGs.

RO6 Validate such framework with real-world Dementia research data from several angles, specifically emphasizing modeling and querying capabilities and state-of-the-art ML approaches.

1.4 Research plan

We introduce a specific research plan that serves as a concrete roadmap to achieve the ROs and, ultimately, materialize the main aim of this work described above. The plan is designed to be modular and iterative, allowing for adaptability in response to emerging findings and technological advancements. It encompasses a range of activities, from the initial investigation of methodological frameworks and terminological foundations to the eventual evaluation of real-world data. Due to the interconnected nature of the objectives, individual components of the plan often address multiple facets that span different research objectives, ensuring a unified yet thorough exploration of the topic. The primary tasks of the plan are as follows.

T1 Design a modular, scalable methodological framework to approach the problem of terminology selection and alignment to integrate semantic technologies into existing neuroimaging and biobanking systems.

1. Follow a layered architecture to incorporate schemas, ontologies, and services.
2. Enable solid and straightforward means for integrating research data management platforms into extract-transform-load (ETL) pipelines for semantic research data integration.

T2 Create ontological extensions specifically crafted to facilitate detailed Dementia research data analysis.

1. Identify gaps in the existing terminology and propose extensions to address them, and reusing relevant classes when necessary.
2. Follow ontology design patterns for newly defined terms.

3. Align the proposed extensions with existing ontologies and upper models.
4. Employ modular building methods to streamline the ontology implementation life cycle.

T3 Develop a low-code transformer module to simplify the data integration process, making the framework accessible to researchers with varying levels of expertise.

1. Implement transformations following the proposed design patterns.
2. Introduce a data descriptor template to couple with usual research outcome data artifacts.

T4 Implement a flexible graph builder able to obtain, transform, and merge sparse ontology, knowledge, and data annotation sources while providing a friendly means for extension and customization.

1. Ensure consistent knowledge merging by employing consistent ID mechanisms.
2. Provide platform-independent KG serialization formats.
3. Employ configuration files to guide the different building steps to limit the need for extra development efforts.

T5 Evaluate the applicability of Graph Database Management Systems (GDBMSs) in Biomedical Research and the context of the framework.

1. Conduct a comprehensive assessment of existing GDBMSs, comparing them with Relational Database Management Systems (RDBMSs) and other NoSQL engines regarding scalability, query languages, and efficiency.
2. Identify scenarios within the biomedical domain where GDBMSs offer advantages, focusing on datasets with complex relationships.
3. Assess the performance of GDBMSs in relationship-centric searches, such as path traversals, and compare it with other database systems.
4. Examine the evolution of GDBMS technology to identify its readiness for deployment in both small prototypes and large, production-ready projects.

T6 Evaluate the framework with real-world Dementia research data.

1. Investigate the effectiveness of the different design patterns.
2. Use GDBMSs to execute different querying scenarios involving biological data at different levels of complexity.

3. Use currently available graph processing libraries to compute metrics and embedding methods and apply them to answer potential questions of interest in the field of Dementia research.

We address all tasks and therefore objectives in a series of three published articles.

In the first article, "*Extending XNAT Platform with an Incremental Semantic Framework*," [Timón et al., 2017], we present an Incremental Semantic Framework that addresses the challenge of semantically annotating research data across three levels of abstraction while retaining pertinent metadata in research data management systems. The framework is applied to the XNAT neuroimaging platform through multiple use cases. This article fulfills the initial task T1 and part of T6 in the research plan, targeting RO1 and partially RO6. It concentrates on the methodology for ontology design and the foundational aspects of research data instantiation.

The second article, [Timón-Reina et al., 2021], provides a narrative review of the application of GDBMSs in biomedical data. It explores the different available GDBMS technologies, assesses their performance, and examines how the research community utilizes them. This article covers RO5 and T5 of the research plan.

In the third and final article of this thesis, [Timón-Reina et al., 2023], we present DemKG as the culminating result of the research. The article outlines the methodology employed, sources of knowledge, modular component implementation, and use cases for validation. This contribution fulfills tasks T2, T3, T4, and T6 in the research plan, addressing RO2, RO3, RO4, and RO6.

Chapter 2

Methods

This thesis is framed at the intersection of several domains, particularly within Knowledge Representation (KR), its storage, the modeling and transformation into KGs, and their application in Dementia research. Thus, our work engages with and extends real-world Dementia research projects, key research areas, and technologies. This section discusses these materials and our methodological approach.

First, we detail the research data sources that underpin the primary motivation of this thesis, providing an overview of their origins and the data modalities that serve as inputs for the following modeling. The subsequent section elaborates on the data capture strategies and the data management system that serves as the central repository for later research stages. We then describe the open-source tools integral to the framework's implementation. Following this, we discuss the methodological approach underlying our ontology extensions and then present the final semantic model. The section concludes with an overview of the software implementations that constitute the full functionality of the DemKG framework.

2.1 Source study research data

This research is conducted in collaboration with the Neurology department at Akershus University Hospital and aligns with the objectives of the Dementia Disease Initiation (DDI) and Precision Medicine Interventions in Alzheimer's Disease (PMI-AD) projects [Fladby et al., 2017]. The DDI study, a longitudinal observational initiative involving all Norwegian health regions and university hospitals, aims to identify early biomarkers for individuals at risk of developing Dementia. The cohort comprises individuals who have either self-reported cognitive decline or have been referred from memory clinics, along with healthy controls sourced from spouses and family of patients. Data collection involves both clinical and biomarker assessments. Clinical data is captured using a Case Report Form (CRF) that includes medical history, physical and neurological examinations, cognitive assessments,

and the Geriatric Depression Score (GDS) [Mitchell et al., 2010]. The cognitive assessments encompass the Mini-Mental State Examination (MMSE-NR) [Kurlowicz and Wallace, 1999], clock drawing test [Mainland and Shulman, 2017], the Consortium to Establish a Registry for Alzheimer's Disease (CERAD) word list test [Fillenbaum and Mohs, 2023], VOSP silhouettes [Quental et al., 2013], psychomotor speed, Trail making A and B [Bowie and Harvey, 2006], and Controlled Oral Word Association Test (COWAT) [Benton et al., 1994]. Biomarker data is obtained from cerebrospinal fluid and blood samples, yielding measurements of proteins, enzymes, and cells pertinent to biological functions like amyloid metabolism, innate immune response, and synaptic activity. MRI scans are conducted for all subjects, and when available, FDG-PET and amyloid PET scans are also performed. The amassed data necessitates robust storage solutions for efficient management, querying, and retrieval.

These studies serve as both the conceptual foundation for the developments proposed in this thesis and the empirical context for validating the stated hypothesis and objectives

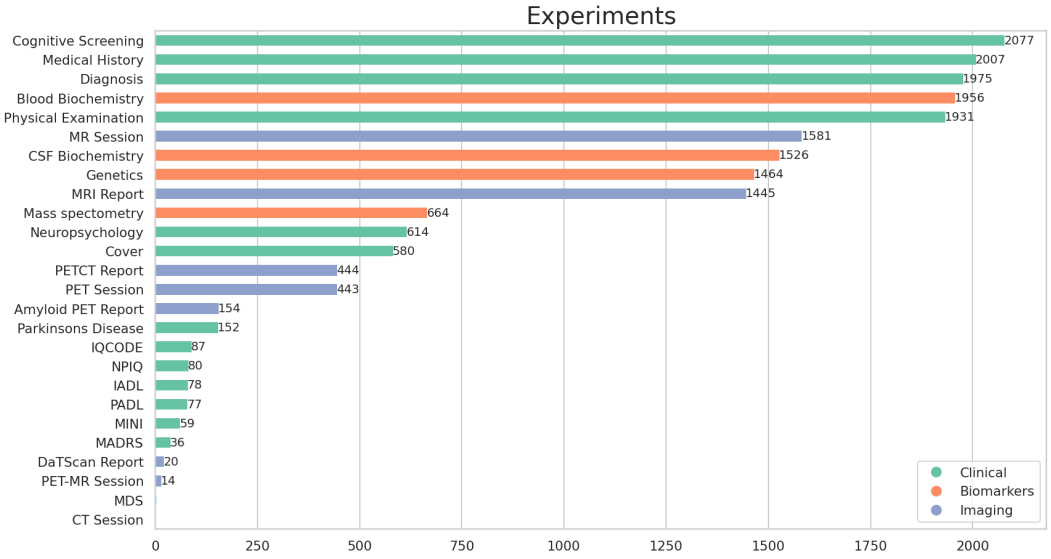


Figure 2.1: Overview of the DDI experimental categories.

2.2 Research data capture and management

To manage diverse datasets and imaging modalities effectively, the data management system needed to fulfill specific requirements and offer key functionalities. These included an open-source license, the ability to store various neuroimaging modalities, customizable data structures for different types of non-imaging experiments, remote access with role-based user administration, and integration capabilities with computational methods. After a thorough evaluation of available options, we selected the eXtensible Neuroimaging Archive Toolkit (XNAT).

XNAT is an open-source software platform designed to facilitate the comprehensive management of neuroimaging and related data. Developed primarily in Java, XNAT provides a robust framework for secure data storage, retrieval, and sharing while also offering customizable data structures to accommodate various research needs. It supports a wide range of imaging data types, including MRI, PET, and CT scans, among others. One of its key features is the ability to integrate with existing data processing pipelines, thereby streamlining the workflow from data acquisition to analysis. Additionally, XNAT's RESTful API enables seamless interoperability with other software tools and platforms, making it a versatile choice for multi-center studies and collaborative research endeavors. Its modular architecture and extensibility make it particularly well-suited for complex neuroimaging projects that require a high degree of customization and scalability. Thus, the platform not only provides robust storage and management features for imaging data but also offers a flexible plugin framework to tailor the system to the study's specific needs.

Our first article outlines how to exploit these capabilities in an ETL pipeline that implements a proposed Incremental Semantic Framework (ISF), elevating raw research data that follow several low-level schemas into formal semantics of a higher level. This integration enables several benefits, such as quality control, as evidenced in the article.

Lastly, given the sensitive nature of the clinical data involved, a secure deployment environment was imperative. The project set up the system within the service for sensitive data (Tjeneste for Sensitive Data, TSD), operated by the University of Oslo. This environment is equipped with stringent access control measures to ensure data security and restrict external internet access.

2.3 Knowledge Graph tooling

In order to manipulate, create, and further process or analyze graphs, we take advantage of many graph-oriented tools.

Although not explicitly designed for graph structures, certain Semantic Web standards and technologies inherently support them. Specifically, RDF offers various mechanisms for interacting with RDF Graphs, and its databases and SPARQL[1] query language are essential for effective utilization. We employed these technologies with ontology development tools for multiple tasks, ranging from ontology construction to enabling graph-based functionalities.

The field of Knowledge Graphs is experiencing significant growth. This expansion has led to the development of a myriad of software and tools designed to address the unique challenges posed by knowledge representation and graph-specific requirements. In the biomedical field, initiatives such as the Knowledge Graph Hub (KG-Hub) [Caufield

[1] https://www.w3.org/TR/rdf-sparql-query/

et al., 2023], Monarch [Mungall et al., 2017], and the “universal biomedical data translator” program from the National Center for Advancing Translational Sciences (NCATS) [Fecho et al., 2022] are instrumental in fostering the development of open-source tools that serve the KG community. These initiatives contribute to the standardization, scalability, and utility of KG tooling, thereby accelerating research and application in various scenarios.

2.3.1 KGX

One of the fundamental tools exploited in this work is the Knowledge Graph eXchange (KGX). KGX is both a serialization of Biolink Model compliant knowledge graphs, and a graph processing library. The KGX format specification is designed as flat files that can be processed, subsetted, and exchanged easily. Each node (or edge) is represented with all of its properties that further describe the node (or edge). The default serialization is TSV. As a library, KGX supports the transformation of graphs into standard formats like KGX TSV, JSON, or RDF, making it easier to integrate disparate data sources. KGX also provides functionalities for merging multiple KGs, thereby enabling the creation of more comprehensive and interconnected knowledge networks.

2.3.2 KG-Hub

KG-Hub is an initiative for standardizing the construction, exchange, and reuse of KGs. The platform employs a modular extract-transform-load workflow to generate graphs aligned with the Biolink Model. It facilitates seamless integration of ontologies from the OBO initiative and offers features such as cached downloads of upstream data, version-controlled builds with stable URLs, and cloud-based, web-accessible storage of KG artifacts. KG-Hub is versatile, supporting a variety of research projects that range from COVID-19 and drug repurposing to microbial-environmental interactions and rare disease studies. Additionally, the platform is equipped with analytical tools for KG manipulation and is closely integrated with machine learning tools for automated graph-based tasks, including node embeddings, link prediction, and node classification.

2.3.3 KG-OBO

KG-OBO focuses on the conversion of OBO ontologies into KGs. It provides a set of tools and guidelines for transforming OBO ontologies into RDF triples, thereby enabling their integration into broader KGs. This is particularly useful in biomedical research where OBO ontologies like Gene Ontology (GO) [The Gene Ontology Consortium et al., 2023] and Human Phenotype Ontology (HPO) [Köhler et al., 2021] are widely used.

2.3.4 High performant graph processing libraries

Executing standard graph operations on large KGs—including filtering, merging, centrality computation, embedding generation, and other machine learning-oriented tasks—poses computational challenges. To address these issues, there are actively developed libraries designed to meet these specific needs. The Knowledge Graph Toolkit (KGTK) [Ilievski et al., 2020] and Graph Representation leArning, Predictions, and Evaluation (GRAPE) [Cappelletti et al., 2023] library are two relevant examples that we employed for computing several graph metrics and analytics.

2.4 Ontology engineering, technologies, and development

We aimed to establish a robust methodology for the design, development, and maintenance of the framework's terminological component. Accordingly, we selected technologies and patterns well-suited for ontology development.

Among the various technologies for ontology development and serialization, we chose to use Semantic Web ontology standards as defined by W3C, given their widespread adoption. Specifically, we employed OWL for ontology definition and RDF for intermediate serialization. OWL's compatibility with Description Logics (DL) [Baader et al., 2008, Krötzsch et al., 2014] allowed us to create logically grounded definitions. To ensure the integrity of the ontology, we used the ELK reasoner [Kazakov et al., 2014] for logical reasoning, which helps in identifying and avoiding errors and inconsistencies early on in the process.

For term definition, we adhered to a systematic approach, following OBO guidelines and utilizing ontology design patterns. We particularly relied on the framework provided by Dead Simple OWL Design Patterns (DOS-DP) [Osumi-Sutherland et al., 2017], which is effective in minimizing errors when defining multiple interrelated terms.

To streamline the building and releasing process, we employed the Ontology Development Kit (ODK) [Matentzoglu et al., 2022]. This tool automated and validated several crucial intermediate steps, including ontology imports and term pattern materialization, thereby facilitating the final ontology-building process.

These topics are described and discussed in articles 1 [Timón et al., 2017] and 3 [Timón-Reina et al., 2023] of this thesis.

2.5 Terminological and semantic model

A primary research objective of this thesis, RO1 with tasks T1 and T2, is to develop a robust semantic model that articulates clear definitions for pertinent concepts within the target

knowledge domain and establishes expressive relationships among them. Additionally, the model aims to facilitate the transition from raw, low-level data to enriched semantics at both the domain-specific and upper-levels of abstraction, emphasizing retaining references to the source, provenance, and other metadata, which is extremely important for tasks like Quality Control.

For this purpose, we introduce the mentioned Integrated Semantic Framework (ISF), detailed in this thesis's first article. The ISF serves as the foundational element for the semantic modeling carried out by the DemKG framework. It offers a structured blueprint for transforming raw research data models into more semantically enriched representations facilitated by domain-specific ontologies.

This framework may process raw data in XML format, adhering to a predefined set of supported XML Schema Definition (XSD) schemas and from source tabular data in CSV or TSV formats. To enhance semantic richness, bio-ontologies supply the requisite terminology and logical axioms defining them. Due to the complexity of the biological domain, the development of these ontologies often targets specific subdomains, leading to a diverse array of reference or domain ontologies.

2.5.1 Bio-ontologies

The fundamental principle in the framework for the foundational semantic model is the utilization of domain reference ontologies to ensure the following aspects:

1. The concept definitions are concise, accurate, and relevant;
2. There exists an active community maintaining the ontology updated;
3. They are widely recognized, cross-referenced, and follow consistent design patterns.

The selection criteria align with the guiding principles of the OBO Foundry. OBO supports a wide array of domain-specific ontologies characterized by clearly defined scopes, concept reusability across different ontologies, and alignment with a unified top-level ontology, namely the Basic Formal Ontology (BFO) [Arp and Smith, 2008]. Relationships within these ontologies are further standardized through the Relations Ontology. Due to these features, OBO ontologies received priority consideration in our selection process.

2.5.2 Biolink

Even employing domain and upper-level ontologies, there are situations where these do not cover all alignment needs, such as with non-OBO ontologies. The Biolink Model is an open-source, universal data model for organizing data in biomedical KGs that defines

entities and the relationships between these entities within translational science. The model serves both as a map for bringing together data from different sources under one unified model, and as a bridge between ontological domains. Biolink provides high-level biomedical concepts (e.g., Protein, Gene, or Disease) and relationships (e.g., "has part", "expressed in"), acting as a gluing component at the higher level of abstraction, completing the design of the semantic model for DemKG.

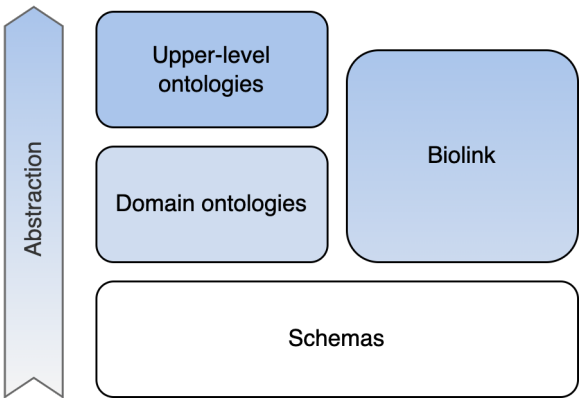


Figure 2.2: Conceptual incremental level of abstraction across the terminological components defined by the ISF.

2.6 Software implementations

In this section, we delve into the software implementations that constitute the DemKG framework. The aim is to provide a clear overview of the technical architecture, the technologies employed, and the rationale behind key design decisions. We will discuss the various modules that make up DemKG, their interdependencies, and how they collectively contribute to achieving the framework's objectives.

DemKG is structured around a modular design, with each component assigned a specific role. The final implementation is composed of three primary modules: the extensions ontology builder, the transformation module, and the KG builder module. This modular approach allows users to integrate or modify individual components as needed or utilize the complete KG construction pipeline.

Each module serves a distinct function and is publicly available in its own GitHub repository. The design, implementation, and validation of the complete DemKG framework are detailed in the third article of this thesis [Timón-Reina et al., 2023], chapter 5.

2.6.1 Extensions ontology builder

Title: DemKG extensions ontology
GitHub URL: https://github.com/demkg-framework/extensions-ontology
License: MIT License
DOI: 10.5281/zenodo.8412054

The extensions ontology builder module generates the terminological and axiomatic extensions in the form of a final OWL ontology. The key gaps addressed include:

- **Phenotypic and Physiological Normal Concepts**: To model phenotypic normality, we employ a DOS-DP pattern that creates the phenotypic normality branch by adapting "normality terms" from "abnormal terms" in the Human Phenotype Ontology and following the Quality-Entity pattern and creating.
- **Abnormal Biomarker Phenotypes**: Given the increasing focus on biomarker-based research in Dementia, we introduce terms not yet defined in domain ontologies that offer detailed logical descriptions linking biological entities, such as proteins and cells, to quantities and anatomical locations.
- **AT(N) Biomarker Profiles**: We extend the Ontology for Biomedical Investigations (OBI) [Bandrowski et al., 2016] classes to include the AT(N) classification system developed by the NIA-AA Research Framework [Jack Jr. et al., 2018], which assesses biological states based on Beta-amyloid deposition, pathologic tau, and neurodegeneration.
- **Assay-Related Classes**: Additional assay and platform definitions missing from OBI are also implemented.

The module is built as an ODK project and integrates several functionalities:

- Importing domain ontology dependencies.
- Applying DOS-DP patterns.
- Manually editing non-systematic terms.
- Materializing implicit relations, particularly class equivalance restrictions, that do not directly translate to graph edges.

In line with modularity and concept reuse principles [Kazakov, 2008], the extensions ontology places new terms as subclasses of relevant parent terms from reference ontologies. This is achieved through ontology extraction and importing pipelines configured in ODK and based on ROBOT commands [Jackson et al., 2019]. The approach allows for the integration

of essential logical axioms while managing ontology size and complexity. Most classes in the extensions ontology are systematically defined using DOS-DP patterns.

The final building step involves executing the relation-graph datalog to materialize subclass and class equivalence restrictions that do not directly translate to graph edges. This is crucial for preserving knowledge in the form of relationships between defined classes.

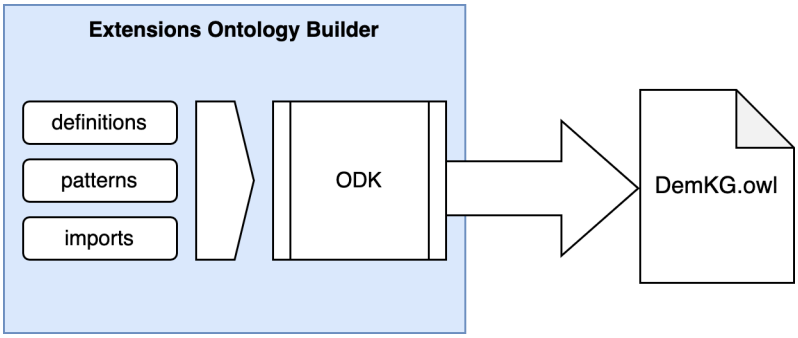


Figure 2.3: An overview of the extensions ontology builder processing.

2.6.2 Research data transformation module

Title: DemKG KG-Transform
GitHub URL: https://github.com/demkg-framework/kg-transform
License: MIT License
DOI: 10.5281/zenodo.8412821

The transformation module delivers a low-code Python package to process source research data into an instantiation graph that aligns with the semantic framework's terminology. The transformation leverages ontology design patterns, focusing on various aspects of scientific data relevant to biomedical and Dementia research. These patterns are mainly rooted in the Ontology for General Medical Science (OGMS) [Scheuermann et al., 2009] and the experimental patterns from OBI. Key features include:

- Support for modeling longitudinal study visits.
- Detailed experimental measurements that capture essential metadata and provenance, and establish explicit links between actors and biological entities involved in processes such as lab assays, omics assays, imaging, cognitive testing, and final diagnoses.

The low-code feature is facilitated through a human-friendly YAML schema, which outlines the mapping rules for generating graph representations and associating relevant ontology classes based on the input dataset's columns and values. A template schema is provided to ease adoption.

The design patterns address critical data capture elements in Dementia studies, including:

- Subject demographics.
- Medical and clinical history.
- Physical examinations.
- Cognitive screening.
- Diagnosis.
- Specimen assays, encompassing laboratory, proteomics, and genomics tests.
- Imaging-derived analyses.

This methodology effectively enables expressive phenotyping, diagnoses, and conclusions based on the experimental evidence contained in the data. Furthermore, the dataset descriptor schema allows for expressing phenotype findings and conclusions based on data with categorical, cutoff, or range assignments. The module is versatile, functioning either as a Command-Line Interface (CLI) program or as a Python module.

```yaml
specimen_assays:
  -
    specimen_assay:
      id_col: csf_id
      date_col: csf_date
      specimen: UBERON:0001359 # 'cerebrospinal fluid'
      measurements:
        -
          assay_type: OBI:2100266 # 'cerebrospinal fluid protein assay'
          date_col: elisa_1_date # The column with the date of the measurement assay
          col_name: csf_ab42 # The name of the column containing the measurement value
          target_analyte: PR:000050063 #'beta-amyloid protein 42 (human)'
          units: UO:0010070 # 'picogram per milliliter'
          value_phenotype_mappings:
            type: cutoff
              mappings:
                '<1051': HP:0025683 # 'Abnormal amyloid beta 42 peptide CSF concentration'
                '>1050': DEMKG:0000150 # 'Normal amyloid beta peptide CSF concentration'
```

Figure 2.4: A snapshot of an specimen assay descriptor.

2.6.3 KG builder module

Title: DemKG KG-Builder
GitHub URL: https://github.com/demkg-framework/kg-builder
License: MIT License
DOI: 10.5281/zenodo.8412831

The KG builder module is implemented in Python and utilizes various KG tooling technologies, supported by initiatives such as KG-Hub, Monarch, and NCATS. Similar to the transformation module, it can operate either as a CLI program or as a Python module. The module orchestrates the creation of the final KG through a series of steps:

- Downloading predefined knowledge sources.
- Transforming sources to KGX format for ingestion.
- Merging KG artifacts to produce the final KG.

Each step is highly configurable, allowing customization through YAML files and transformation code when specific knowledge sources require it. The module can function as a complete building pipeline or execute individual steps as needed.

Download

The download step fetches the necessary knowledge sources for constructing the final DemKG artifact. While it is configured to obtain default DemKG sources, it is possible to further customize or extend through a YAML file. This file allows users to specify source URLs and optional local names to prevent file naming conflicts. Users can also choose whether to cache existing files or overwrite them, an essential feature given the potentially large size of some knowledge sources.

Transform

The transform step ensures that all sources are available in KGX format for the final merge. Given that transformation requirements can vary significantly between sources, this step is more variable and may require additional development. The module designates a specific location for users to place their transformation code. Although the transformation operation is user-defined, the KG builder expects the output to be in KGX format.

Merge

The merge step primarily utilizes KGX merge functionality to identify, combine, and serialize all nodes and edges from various sources into the final KGX graph. This step is also configurable through a dedicated YAML file, which includes all default DemKG sources and can be easily customized or extended. Additional graph operations, such as computing structural statistics, and other serialization formats like RDF in n-triples, can also be specified.

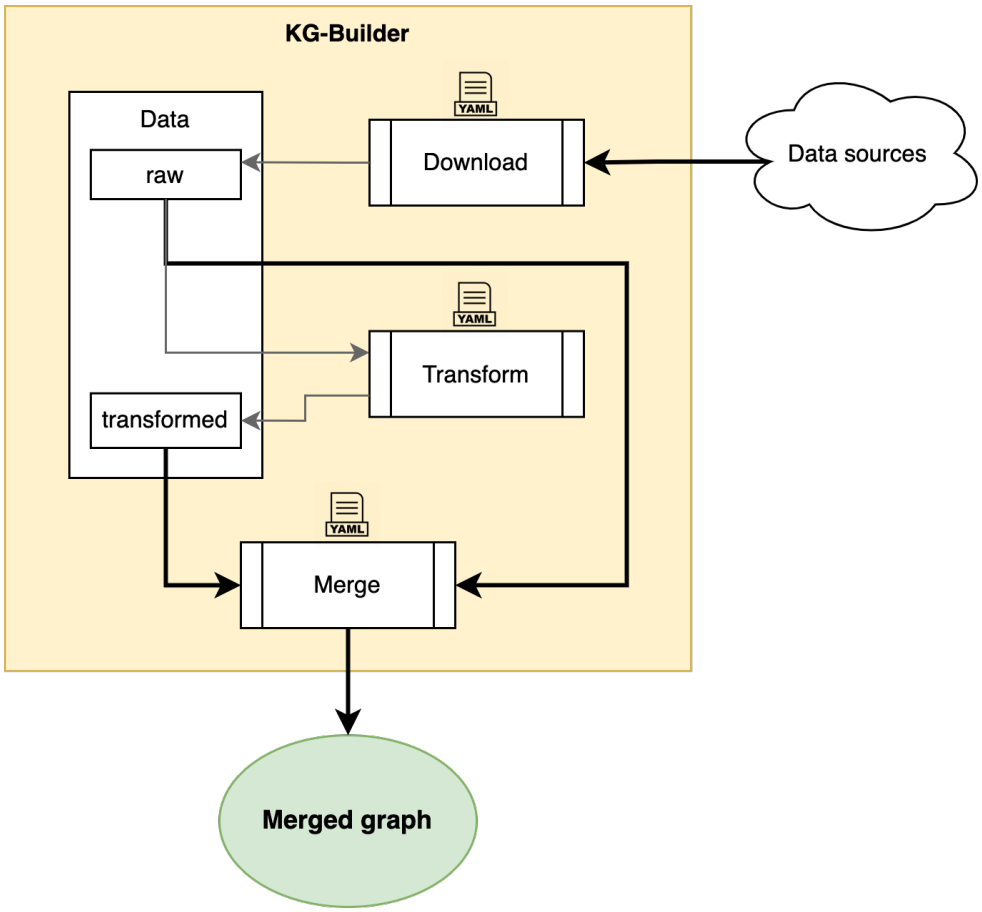


Figure 2.5: Overview of the KG builder elements and processing flow.

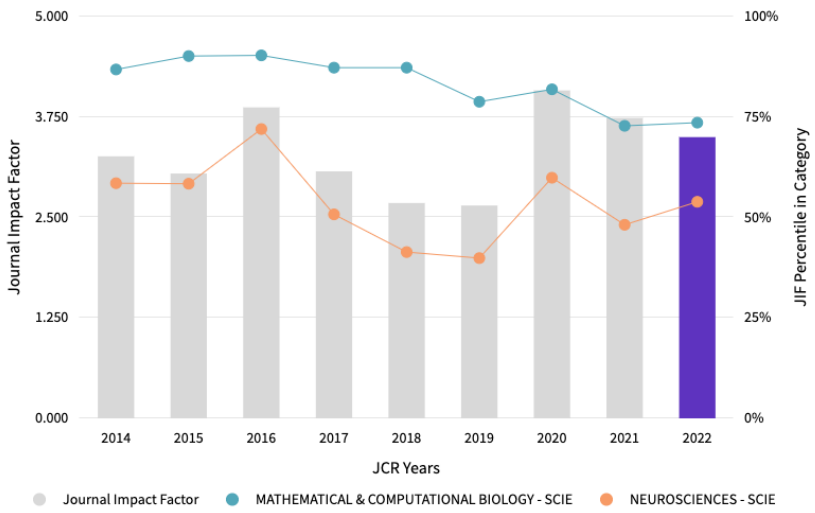
By following these design principles, the KG builder module provides a flexible and efficient means to construct knowledge graphs tailored to specific research needs by offering a modular and configurable approach.

Chapter 3

Extending XNAT Platform with an Incremental Semantic Framework

Title	*Extending XNAT Platform with an Incremental Semantic Framework*
Journal	Frontiers in Neuroinformatics
Authors	Santiago Timón, Mariano Rincón, and Rafael Martínez Tomás
Published	31 August 2017
Impact Factor	3.074
JCR Quartile	Q1
DOI	10.3389/fninf.2017.00057

3.5
2022 Journal Impact Factor



frontiers in Neuroinformatics

ORIGINAL RESEARCH
published: 31 August 2017
doi: 10.3389/fninf.2017.00057

Extending XNAT Platform with an Incremental Semantic Framework

Santiago Timón[1,2,3]*, *Mariano Rincón*[1] and *Rafael Martínez-Tomás*[1]

[1] Departamento de Inteligencia Artificial, Universidad Nacional de Educación a Distancia, Madrid, Spain, [2] Department of Neurology, Akershus University Hospital, Lørenskog, Norway, [3] Intervention Centre, Oslo University Hospital, Oslo, Norway

Informatics increases the yield from neuroscience due to improved data. Data sharing and accessibility enable joint efforts between different research groups, as well as replication studies, pivotal for progress in the field. Research data archiving solutions are evolving rapidly to address these necessities, however, distributed data integration is still difficult because of the need of explicit agreements for disparate data models. To address these problems, ontologies are widely used in biomedical research to obtain common vocabularies and logical descriptions, but its application may suffer from scalability issues, domain bias, and loss of low-level data access. With the aim of improving the application of semantic models in biobanking systems, an incremental semantic framework that takes advantage of the latest advances in biomedical ontologies and the XNAT platform is designed and implemented. We follow a layered architecture that allows the alignment of multi-domain biomedical ontologies to manage data at different levels of abstraction. To illustrate this approach, the development is integrated in the JPND (EU Joint Program for Neurodegenerative Disease) APGeM project, focused on finding early biomarkers for Alzheimer's and other dementia related diseases.

Keywords: biomedical ontologies, Semantic Web, knowledge management, XNAT, data exchange, data analysis, Neurodegenerative Diseases

OPEN ACCESS

Edited by:
Jose Manuel Ferrandez, Universidad Politécnica de Cartagena, Spain

Reviewed by:
Jose M. Molina, Universidad Carlos III de Madrid, Spain
Fernando Perez, University of California, Berkeley, United States

***Correspondence:**
Santiago Timón
santiago.timon.r@gmail.com

Received: *02 May 2017*
Accepted: *14 August 2017*
Published: *31 August 2017*

Citation:
Timón S, Rincón M and Martínez-Tomás R (2017) Extending XNAT Platform with an Incremental Semantic Framework. Front. Neuroinform. 11:57. doi: 10.3389/fninf.2017.00057

INTRODUCTION

Nowadays, neuroscience research projects take place in multidisciplinary, heterogeneous multi-center environments, where an efficient mean of data exchange is crucial. One of the main challenges is the accurate and effective exchange of data for its subsequent analysis, that leads to the need of a common structure, data standardization or some mediation strategies (Ashish et al., 2010). Some currently in use archiving solutions, as reviewed in Izzo (2016), are the Extensible Neuroimaging Archive Toolkit (XNAT) (Marcus et al., 2007), the Collaborative Informatics and Neuroimaging Suite (COINS) project (Scott et al., 2011), or the eXTENsible platform for biomedical Science (XTENS) (Corradi et al., 2009).

Despite the flexibility and ease of customization offered by the mentioned archiving systems, data scalability is somehow limited, as significant changes in the data model typically require fine configuration of the database or an important reorganization. These shortcomings have been addressed by the use of ontologies and Semantic Web technologies (mainly OWL[1], RDF[2], and SPARQL[3]) (Hoehndorf et al., 2015). The Mayo Clinic made one of the first examples of such approach by applying Linked Data principles to its Electronic Health Records (Pathak et al., 2012a).

[1]https://www.w3.org/TR/owl2-primer/
[2]https://www.w3.org/TR/rdf11-primer/
[3]https://www.w3.org/TR/sparql11-overview/

They leveraged publicly available data from the Linked Open Drug Data cloud (Samwald et al., 2011) to federated querying for type 2 diabetes patients. Following the same principle, Leroux and Lefort (2015) showed an efficient approach to enrich the semantics in clinical trials. They developed a semantic, linked data model from CDISC Operational Data Model[4], focusing just on the easy data sharing and consumption, and leaving further modeling and reasoning for the future. On a more domain-specific context, Hsu et al. designed an ontology-driven system employing an application ontology that imports and aligns ontologies from different domains (Hsu et al., 2015). It integrates phenotypes generated through analyses of available clinical data sources. Their approach demonstrated how an ontological framework could help to enforce consistent data representation and even enable further studies to identify clinical predictors. Also, numerous approaches have been proposed for complex knowledge intensive tasks in the past years, like radiological assistance (Mejino et al., 2008), surgical planning (Mechouche et al., 2007, 2009), or clinical management (Sonntag, 2008) and patient care systems (Su and Peng, 2012).

Notwithstanding the obvious growth in its application, the adoption of ontological frameworks shows some drawbacks and is still a challenging and time consuming venture (Hastings et al., 2014). There exists a trade-off between the language expressiveness and its computational tractability that requires making decisions about the necessary level of description. Usually, the use of highly descriptive ontologies alone results in *ad-hoc* implementations for domain-specific solutions with poor scalability that complicates raw data extraction for less knowledge-aware tasks. Furthermore, ontology selection, alignment, and mapping require the collaboration of domain experts and development staff, in addition to the steep learning curve for new users of ontologies. Ontology engineering methodologies, such as the NeOn Methodology (Suárez-Figueroa et al., 2012) provide a methodological guide for addressing several of the mentioned issues, usually targeted at a final high-level ontological ecosystem. However, leaving behind intermediate low-level data is problematic when the goal is integrating complex, distributed systems. The loss of the original data structure compromises data quality and limits the possibilities for its manipulation at the same time. A Bottom-up approach that supports all description levels simultaneously is more convenient for these projects. It has been successfully applied in other domains, for e.g., in the video analysis domain (Duan et al., 2003).

In this article, we describe an incremental semantic framework; a methodological approach to address the problem of enabling semantic-based modeling in already implemented research archiving systems. Consequently improving data management, from low-level data to semantic and logical concepts. Built with Semantic Web technologies and using biomedical ontologies, the framework provides a model for homogenous data access and reasoning over multi-modal neurological data.

The design of the framework follows a bottom-up, layered approach, allowing working with the data at different levels of description. The framework adds reasoning capabilities from implicit relations and logical definitions to derive new data, as well as to perform data consistency checks for Quality Control (QC). The use of Linked Data principles enables inter-data linking, opening the door to reference external data sets. Also, having a highly linked dataset eases data inspection from different conceptualizations (project, subject, disease, etc.), a highly desirable feature for pattern discovery and studying the relationship between diseases as the dataset grows.

Our proposal differs from previous works in its focus on advanced querying and reasoning without losing low-level data, while taking advantage of already available and widely used archiving platforms. Particularly, we chose XNAT as the backbone for managing clinical and imaging data, for its rich set of features and its flexible and customizable design.

To illustrate the benefits of the framework, this work is encompassed in the JPND (EU Joint Program for Neurodegenerative Disease)[5]/APGeM project[6], aimed at finding early biomarkers for Alzheimer's and other dementia related diseases. It comprises a significant amount of data from different subdomains and modalities, such as neuroimaging, biochemistry, clinical/neuropsychological screenings and genetics, setting up a proper scenario to push and test the framework with a current ongoing neurological research effort.

The remainder of the paper is organized as follows. In Section Material and Methods we describe the design and technological methodology, as well as the data from APGeM's project. Next we exemplify the utility of the framework through various use case applications in Section Results. Finally, in Section Discussion we discuss the benefits, problems encountered and limitations of our implementation and conclude in Section Conclusion.

MATERIALS AND METHODS

This section starts describing the data from the APGeM project. It is part of the driving material and an example of application of the semantic framework. Later, in Section Data Management with XNAT Platform we describe the features of the XNAT platform. In Section Framework Design we outline the decisions made to design each layer of the ontological framework. Finally, in section Data Transformation and Storage, we describe the details of the transformation and loading of the data for persistence.

The related code that is not core to APGeM is available at https://bitbucket.org/apgem-isf/ under Apache Licence, version 2.0.

APGeM Project Data

The APGeM project, where this work is encompassed, is focused on finding early biomarkers for Alzheimer's and other dementia related diseases (Fladby et al., 2017). It comprises individuals assessed with subjective cognitive decline (SCD) (Jessen et al.,

[4]https://www.cdisc.org/standards/transport/odm

[5]http://www.neurodegenerationresearch.eu/

[6]http://www.neurodegenerationresearch.eu/publication/apgem/

2014), mild cognitive impairment (MCI) (Albert et al., 2011), dementia, and healthy controls.

Subjects were recruited from January 2013 to January 2017 and examined following a standardized protocol. Recruitment was based on two main sources: (1) self-referred patients following advertisements in media, newspapers, or news bulletins, and (2) recruited patients among referrals to regional memory clinics. In addition, cognitively healthy controls were also included from spouses of patients with dementia/cognitive disorder, and from patients who completed lumbar puncture for orthopedic surgery. Participants were staged as controls, SCD or MCI using published criteria based on the comprehensive assessment program. Controls were further classified as having normal or abnormal cognitive screening and with or without first-degree relative with dementia.

A case report form (CRF) was developed, comprising medical history (captured from subject and informant separately), and physical and neurological examinations including the 15-item Geriatric Depression Score (Mitchell et al., 2010). The cognitive examination included the Mini Mental State Examination (Folstein et al., 1975), non-verbal cognitive screening (The clock drawing test) (Shulman, 2000), verbal memory (Fillenbaum et al., 2008), visuoperceptual ability, psychomotor speed, and divided attention (Trail making A and B and word fluency). The dataset also included relevant biomarkers for Alzheimer's and other dementia related diseases, obtained from Cerebrospinal fluid and blood samples.

All subjects were referred to a standardized magnetic resonance imaging (MRI) scan protocol; including high resolution structural scans. A sub-set of subjects also underwent an extended MRI protocol including advanced diffusion weighted sequences as well as multiple positron emission tomography (PET) modalities.

Data Management with XNAT Platform

The Extensible Neuroimaging Archive Toolkit (XNAT—RRID:SCR_003048) is an archiving software platform designed to facilitate common management and processing tasks for neuroimaging and related data, providing a secure storage and access layer. XNAT's architecture follows a three-tier design pattern that includes a relational database backend, Java-based middleware engine, and a web-based user interface.

The key of XNAT's flexibility resides in the XML-based data model that defines the data-types that are to be handled by the deployed system. XNAT uses these XML schemas[7] (XSD) to generate custom components, content, and logic for each of the tiers: (1) a relational database structure is generated, equivalent to the elements defined in the XSDs; (2) middleware classes are generated that can be used by developers to implement custom functionality that utilizes the XNAT database; and (3) user interface content, including navigation menus, search options, and data tables. This building mechanism allows research groups to customize data-types and interfaces for storing the relevant data to their studies. The level of this customization is left to developers, going from implementing simple types and questionnaires to complex data structures, interactive interfaces, and business logic.

Another fundamental part is the REST (Fielding and Taylor, 2002) API. It allows interacting with XNAT through HTTP protocol to support basic actions like Create, Read, Update, and Delete resources, as well as more advanced features like data searching and listing, which permits to integrate external pieces of software with XNAT.

Finally, XNAT also ships a pipeline engine that tightly integrates and manages processing pipelines into XNAT's workflow. This was another key feature for the platform selection process, since pipeline execution is critical in Neuroimaging research to develop tasks such as image quality control and automated segmentation.

To this day, there are several publicly available solutions to manage clinical and omics data more efficiently than XNAT, such as BRISK, caTRIP, cBio Cancer Portal, G-DOC, iCOD, iDASH, and tranSMART (Scheufele et al., 2014; Canuel et al., 2015), existing the option to implement a distributed data warehouse system and leave XNAT in charge of neuroimaging data. However, while adapting and customizing XNAT to fit the project needs was a time consuming task, the learning curve was applied only to one system. This allowed for better understanding and, consequently, maximizing the exploitation of XNAT's features.

Framework Design

Conceptually, the framework follows an n-tiered incremental design, composed of three layers, or levels (**Figure 1**): schema, formal and domain. This approach intends to add the complexity cumulatively, in a way that is possible to access low-level data easily (schema and formal levels) and look for further relations and descriptions based on logical axioms at the same time (formal and domain levels). The schemas and ontology acronyms included in **Figure 1** are described in related subsections.

The schema level is the entry point of the framework; it defines the source data structure through XML schemas. The formal level delivers the data modeled with vocabularies under Semantic Web standards. It augments the basic semantics of the Schema level introducing more abstract concepts. These concepts are defined through Description Logics and translated to a RDF graph model without losing completely its source, which allows low-level inspection and data retrieval and also introduces more refined provenance descriptions. Finally, the domain level provides more expressive descriptions to enable further reasoning and query capabilities, for instance, using richer domain specific ontologies to include neuroanatomical terms and mereological axioms.

Schema Level

The core data model of XNAT supports the storage of imaging and custom clinical data, laying the foundation for the schema level, the first layer of the semantic framework. XNAT itself models the basic organizational and imaging data structures, leaving further extensions for other three schemas used in this

[7]https://www.w3.org/TR/xmlschema-0

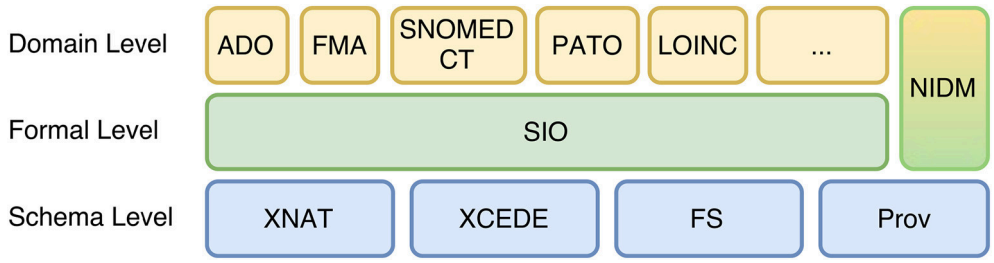


FIGURE 1 | The three component layers of the framework, ordered by abstraction.

layer, XCEDE, FreeSurfer (FS) (FreeSurfer, RRID:SCR_001847) and W3C Provenance data model[8].

While XNAT schema is well fitted for data persistence, its expressivity is somehow limited for describing the study design. We use the XCEDE (XML-based Clinical and Experimental Data Exchange) schema (Gadde et al., 2012) (XCEDE Schema, RRID:SCR_002571) to keep the imaging part of the CRF and describe the study and protocol design under the same specification. The existing overlap between XNAT and XCEDE models facilitates mapping data in both ways and complements the core data model of XNAT.

We leave XNAT schema to focus on data persistence and, as a previous step before introducing more descriptive semantics, employ XCEDE to describe the study protocol in an exchangeable format and link to ontology terms from upper levels in the framework through the "Terminology" component of the schema.

To integrate XCEDE import/export processes properly, we have implemented an XNAT service extension following the same principles as its native REST API to serve study data in XCEDE format. The service serves data by employing several transformation scenarios designed for each resource type defined in the model.

The XNAT community provides the FreeSurfer schema, enabling a means to store FreeSurfer results into XNAT and share them between researchers. Furthermore, having a results XML model eases its processing at higher levels in the framework.

The schema level makes possible to work with XNAT's native data format for low-level data processing, while enabling at the same time data sharing and further modeling through less platform specific schemas. This is very valuable in situations where low-level inspection is needed and abstractions are not beneficial or even counterproductive.

Formal Level

The formal level provides an entry level to model the data through Semantic Web technologies. It serves as the foundational layer to model XNAT experiment data as information entities that describe data, studies and protocols, and which could be further aligned or mapped to specific domain ontologies. It improves low-level semantics by introducing logical definitions with Description Logics (DL), more powerful sharing mechanisms with data linking, query strategies, and finally enabling DL reasoning.

We used NCBO's Bioportal (Musen and Noy, 2011; Whetzel et al., 2011) (BioPortal, RRID:SCR_002713) to find the most suitable ontology. After evaluating various ontologies based on the Basic Formal Ontology[9] (BFO, RRID:SCR_004818) upper-level model, such us the Ontology of Clinical Research (OCRe) (Sim et al., 2014) (Ontology of Clinical Research, RRID:SCR_010392), the Translational Medicine Ontology (TMO) (Luciano et al., 2011), the Semanticscience Integrated Ontology (SIO) (Dumontier et al., 2014) (Semanticscience Integrated Ontology, RRID:SCR_010427), and the Neuroimaging Data Model (NIDM)[10] (Keator et al., 2013) (Neuroimaging Data Model, RRID:SCR_013667), we concluded that SIO covers more terms related to low-level information representation in contrast with OCRe. Also, SIO can be seen as the supported successor of TMO, as it emerged from considerations in the TMO effort. Finally, NIDM is less formal than SIO, but models in more detail concepts related to neuroimaging. On this basis, we decided to employ an alignment of SIO and NIDM as the foundational ontologies to model CRF and imaging data. On the one hand, SIO was used to describe studies and protocols and also to model information entities and experiment data. On the other, NIDM was used to model important provenance and processing neuroimaging results data (Maumet et al., 2016).

At this level, the core elements in the base XNAT data model had to be properly mapped to concepts of SIO. For versions 1.6.x, these elements were Project, Subjects, and Experiments, and some of them lack of direct correspondence with SIO. Most of the mapping process is as detailed bellow.

The term "experiment" in the SIO ontology is defined as an "investigation that has the goal of verifying, falsifying, or establishing the validity of a hypothesis," while for XNAT it is an event by which data is acquired. Therefore, the meaning for "experiment" differs between them and we found "data collection" a suitable entity to model experiment data in XNAT's sense, encoding final literal data with "data item" instances. The description for the entity "data collection" is defined as the process of acquiring information. Adding the insertion/collection date to "data collection" instances complies

[8]https://www.w3.org/TR/prov-dm/

[9]http://ifomis.uni-saarland.de/bfo/

[10]http://nidm.nidash.org/

with XNAT definition of experiment. Hence, the basic starting point to model experiment data is using Data collection class for experiment instances, which has output sub-sections as data set instances. These specify the data fields with has data item property and data item instances. The final values are literals related with has value data type property. Formally in DL notation:

$$Data\ \ collection \sqcap (\exists has\ output.(Data\ set \sqcap (\exists has\ data\ item.(Data\ item)))$$

Figure 2 depicts the basic means to represent an experiment and its data. It is important to note that, depending on the experiment type, the way of obtaining raw values may differ and should be consequently modeled, distinguishing between observations (a doctor's assessment), measurements with values and units (the amount of blood cholesterol) or test outputs (the T-Score for TMT test).

Domain Level

Up to this level, the meaning of the data elements is still kept at low level, leaving the interpretation to *ad-hoc* processes or humans from coding conventions. The purpose of the domain level is to provide high-level semantics and, when possible, logical definitions for the concepts depicted in the data and even rules to further enrich the model. This level tends to be specific to the application or context of the project, thus the ontology selection and modeling decisions depend heavily on it. We demonstrate the building of this level through its application to the Alzheimer's Disease domain.

The Alzheimer's disease ontology (ADO) (Malhotra et al., 2014) (ADO, RRID:SCR_010289) is the first bridge for our use case domain context, focused in Alzheimer's and related diseases. ADO was developed with the purpose of containing information relevant to four main biological views: preclinical, clinical, etiological, and molecular/cellular mechanisms, making possible to map and classify most of the CRF items from APGeM project. The SNOMED CT (Cote and Robboy, 1980) ontology is widely adopted because of its comprehensive clinical terminology. It was used to cover many of the leaf clinical terms in almost every experiment type. To reference anatomical entities we selected the Foundational Model of Anatomy (FMA) (Rosse and Mejino, 2003) (FMA, RRID:SCR_003379) because of its completeness and robust representation of the anatomical reality (Zhang et al., 2003). The Phenotype And Trait Ontology (PATO)[11] was employed to represent biological and phenotypic qualities. The Logical Observation Identifiers Names and Codes (LOINC)[12] (Huff et al., 1998; McDonald et al., 2003) (Logical Observation Identifier Names and Codes, RRID:SCR_010341) was a suitable terminology to map biochemical tests (Bakken et al., 2000), complemented with SNOMED terms. Finally, genetics were mapped to Gene Ontology concepts (Ashburner et al., 2000; Gene Ontology Consortium, 2010). **Table 1** shows a summary of the application of the ontologies to the different sub-domains.

[11] http://obofoundry.org/ontology/pato.html
[12] https://loinc.org/

TABLE 1 | Relation of the component parts of the CRF with subsections and the ontologies with which are modeled.

CRF experiment/ questionnaire	Subsections	Ontology/Vocabulary
Subject demographics		PATO SNOMED CT
Medical history	Social information Family history Current medical history (participant and informant) Current medication Stimulants Other bodily functions Previous medical history Geriatric depression scale	ADO SNOMED CT Disease Ontology
Cognitive screening	MMSE CERAD word list Trail Making Test COWAT (FAS) VOSP silhouettes Clinical dementia rating	ADO SNOMED CT Disease Ontology
Physical examination	General somatic examination Neurological Exam UPDRS Modified UPDRS	ADO SNOMED CT FMA
Diagnosis	Staging Etiology	ADO SNOMED CT Disease Ontology
Biochemistry	Blood tests Spinal puncture (CSF)	ADO LOINC Snomed CT
Genetics		ADO Gene Ontology
Imaging reports		NIDM ADO SNOMED CT FMA

In a typical research project, each experiment type introduces a significant amount of variables (more than 1,100 categorized across several sub-domains in our use case) that need to be mapped to concepts from domain ontologies, implying a very time consuming task. To assist and reduce the time needed in the process of finding term candidates, we developed a script that uses XNAT's search engine through PyXNAT library (Schwartz et al., 2012) (pyxnat, RRID:SCR_002574). For each data-type schema, it inspects complex and simple types to extract the variables to be mapped. Then, for each variable a query is sent to Bioportal's search endpoint with a list of candidate ontologies. The response is a collection of candidate terms for the variable, among other related information, such as the ontology in which the term is defined. The output is an XML file with possible term mappings for each variable. This process has saved a fair amount of time and resources for the ontology and concept selection.

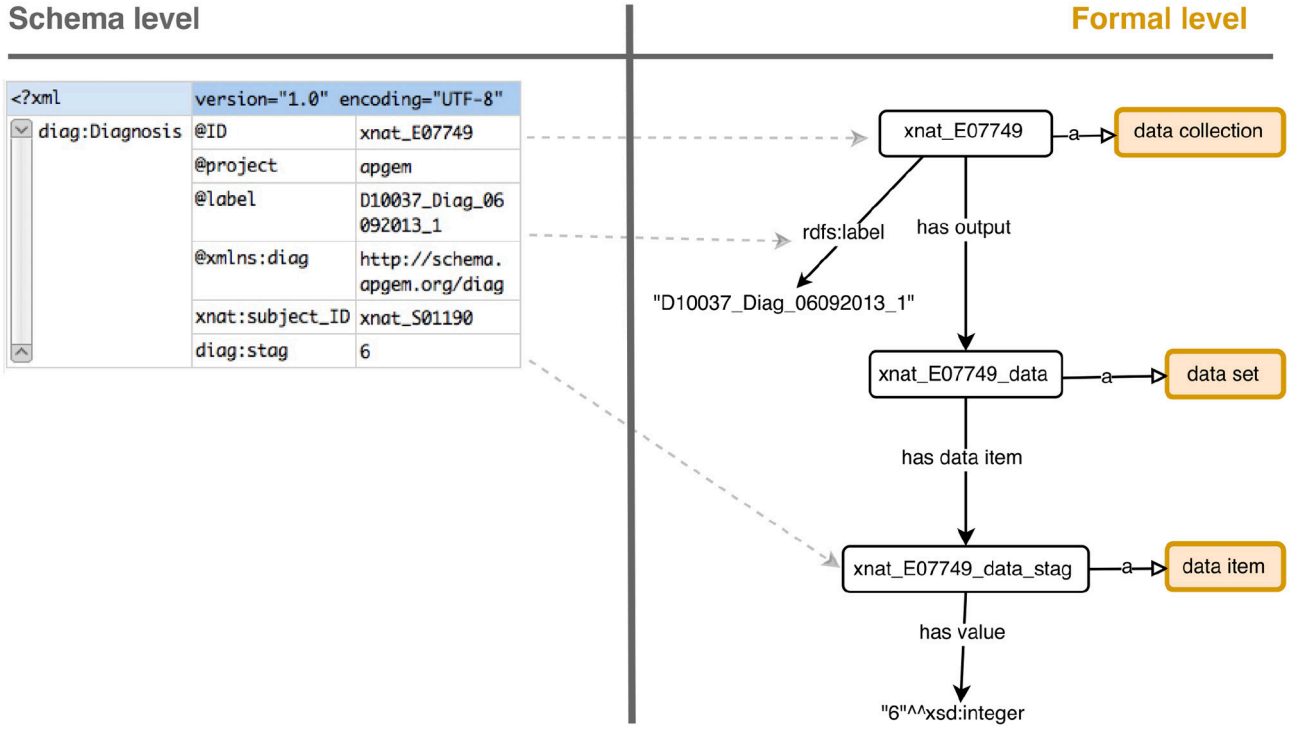


FIGURE 2 | Schema to formal level transformation from a stripped down experiment for exemplification. XNAT XML data is transformed to an RDF graph using SIO classes and properties. Instances are represented as white rectangles and classes as rounded orange rectangles.

Figure 3 shows an example of mappings at formal and domain level.

The domain level for the project was built through the alignment of the selected ontologies. We imported them when possible and, for those too big or broad to be imported, we followed the MIREOT process (Courtot et al., 2011) to include terms of interest. Finally, further logical restrictions and rules relevant to the domain of the use case were defined.

Data Transformation and Storage

At the schema level, the mappings were almost direct between XNAT data model and XCEDE. The transformation was accomplished with XSLT[13] (eXtensible Stylesheet Language Transformations), served on the fly over XNAT's API endpoint. However, before entering the semantic framework, XNAT source data was transformed and mapped to the target model.

To expose subject and experiment data coming from XNAT as RDF, the Extract-Transform-Load (ETL) pipeline depicted in **Figure 4** was implemented.

The workflow is as follows: when any update operation is performed in XNAT the pipeline retrieves the XNAT resource XML and, executes the xnat2RDF script, which transforms it to RDF format using both formal and domain level models. These generated triples are then processed for reasoning, using Pellet reasoner (Sirin et al., 2007) and SPIN[14] (SPARQL Inferencing Notation) API. The output triples from the reasoner script are then loaded into a Jena (JENA: A Semantic Web Framework for Java, RRID:SCR_001766) Fuseki 2[15] triplestore instance.

The primary criterion for the selection of technologies was the ease of integration between the different parts of the workflow, in spite of sacrificing efficiency in some of the steps. Because the execution of this transformation process is made "offline," its performance is not critical to the system's usage. Nevertheless, the execution time is restrained, lasting a couple of seconds per complete subject data (demographics and all experiment data included in the CRF), and less than one second for individual resources.

Fuseki SPARQL Server performs very well in most of the triplestore related operations (Butt and Khan, 2014), although it suffers from write performance problems (Kilintzis and Beredimas, 2014). The reasoning step can be tuned and adapted to use different OWL profiles to reduce execution time. It would be also beneficial to use high-performance reasoning engines like Konclude (Steigmiller et al., 2014), the winner of OWL Reasoner Evaluation 2015 (Parsia et al., 2017). However, these changes would turn into a slightly more complex setup for the ETL process.

We followed the recommendations from the Interoperability Solutions for European Public Administrations (ISA[2]) for the

[13]https://www.w3.org/TR/xslt

[14]http://spinrdf.org/

[15]https://jena.apache.org/documentation/fuseki2

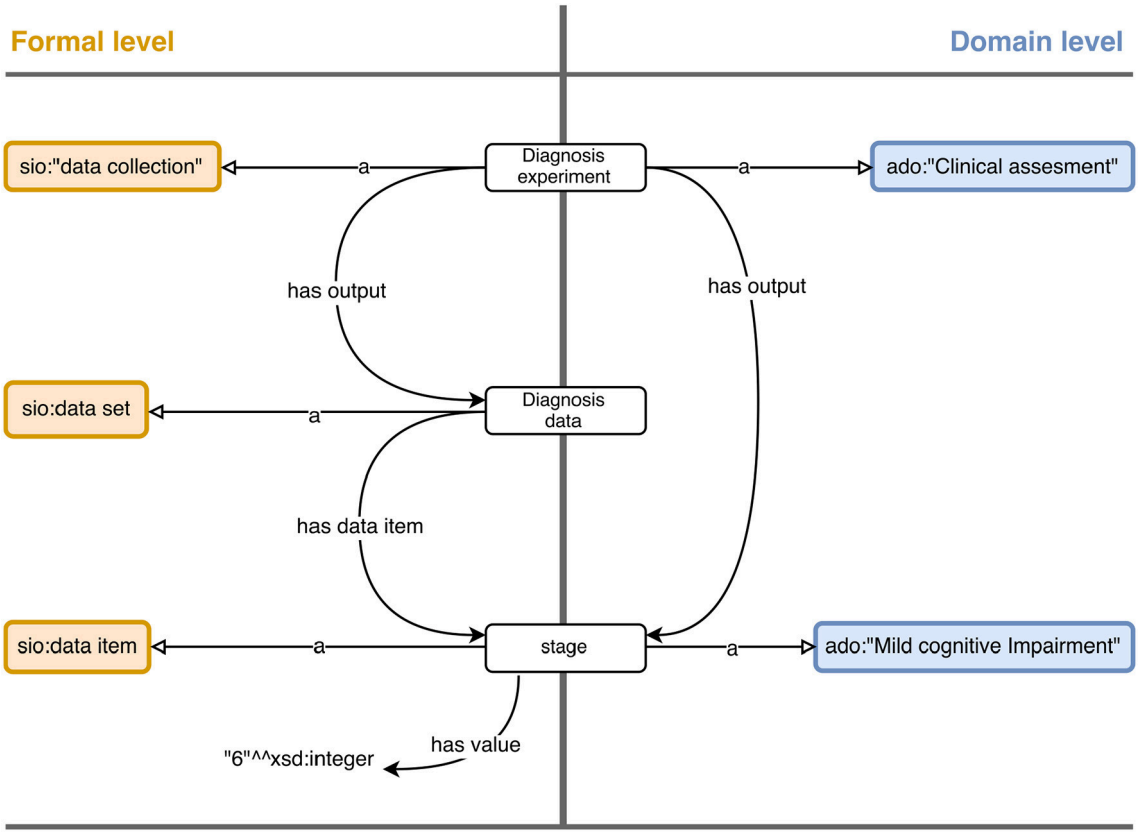


FIGURE 3 | A diagram showing the relations building the assertion that a subject is staged with Mild Cognitive Impairment. The formal level improves the semantics of experiment data, but is still attached to raw values. The domain level introduces specific concepts for a given domain, in this case diagnosis in Alzheimer's Disease. Instances are represented as rounded white rectangles and classes as rounded colored rectangles.

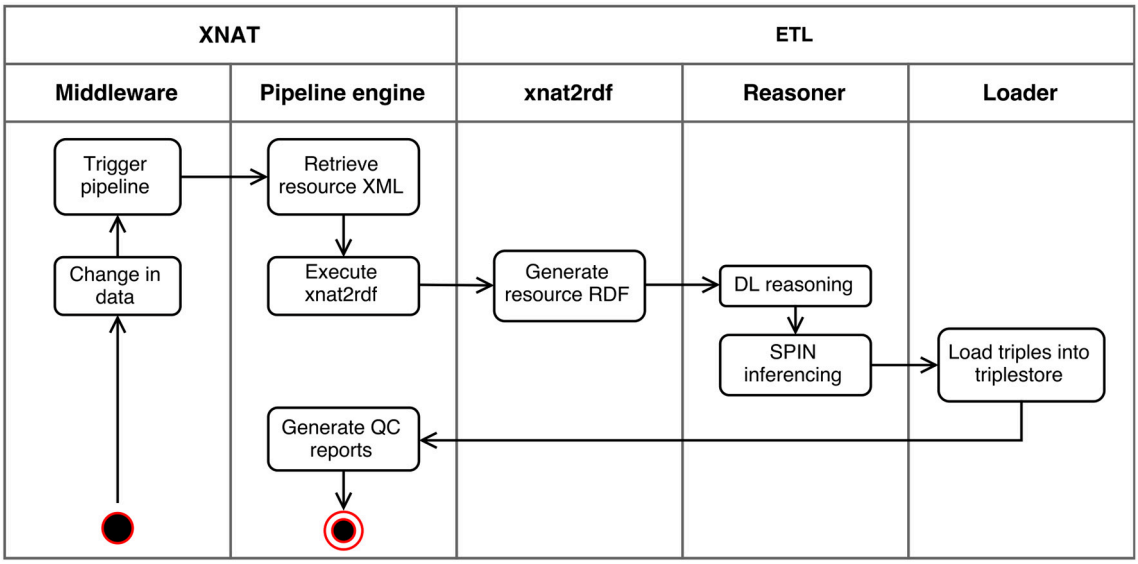


FIGURE 4 | Activity diagram of the ETL pipeline. When any change in the data is registered by XNAT's middleware, the pipeline engine executes the xnat2rdf script passing the XML of the changed resource. This script transforms XNAT XML to RDF, which is processed by the reasoner to execute DL and SPIN inferencing and the resulting triples loaded into the triplestore. Finally QC related data is processed for reporting.

Code 1 | SPIN constraint to determine if a subject meets exclusion criteria.

```
PREFIX sio: <http://semanticscience.org/resource/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX apgem: <http://www.apgem.org/resource/>
PREFIX snomed: <http://purl.bioontology.org/ontology/SNOMEDCT/>

# SPIN reserved word "this" refers to the evaluated instance of
# 'Study subject'
ASK WHERE {
  # This data is obtained from Medical History experiment, previous and
  # current medical history sections.
  ?this sio:SIO_000062 ?mhExperiment.
  ?mhExperiment a apgem:apgem_0003 ; sio:SIO_000312 ?mhdata.

  # The exclusion criteria is met when the subject has filed
  # any of these symptoms:
  # Cerebral infarction, cerebral hemorrhage, epilepsy,
  # head trauma with loss of consciousness,
  # infection in CNS, bipolar disorder, psychosis,
  # delirium/confusion or long term exposure to solvents
  # and malignancy.
  ?mhdata sio:SIO_000028*/sio:SIO_001277 ?cb, ?ch, ?epilepsy, ?ht, ?cnsInfection,
    ?bipolar, ?psychosis, ?delirium, ?exposure.

  # each data item 'denotes' the conditions under SNOMED
  # and the item must have 'true' as value
  ?cb sio:SIO_000020 snomed:432504007 ; sio:SIO_000300 true.
  ?ch sio:SIO_000020 snomed:274100004 ; sio:SIO_000300 true.
  ?epilepsy sio:SIO_000020 snomed:84757009 ; sio:SIO_000300 true.
  ?ht sio:SIO_000020 snomed:82271004 ; sio:SIO_000300 true.
  ?cnsInfection sio:SIO_000020 snomed:128117002 ; sio:SIO_000300 true.
  ?bipolar sio:SIO_000020 snomed:13746004 ; sio:SIO_000300 true.
  ?psychosis sio:SIO_000020 snomed:69322001 ; sio:SIO_000300 true.
  ?delirium sio:SIO_000020 snomed:2776000 ; sio:SIO_000300 true.
}
```

design of persistent URIs[16] that represent the generated resources (instances).

RESULTS

To illustrate the utility of the proposed design methodology, our framework was integrated into the system environment of APGeM. In order to ensure secure access to sensitive medical data, the environment runs on the Services for sensitive data (TSD) provided by the University of Oslo.

The following sections describe how the integration of the framework enabled data science researchers to engage QC, subject classification, and advanced reporting tasks through semantic querying and logical reasoning.

Data Quality Control

Nowadays, the data managed in neuroscience research projects cover very different biomedical fields and is therefore gathered by several, diverse means, such as laboratory reports for biochemical tests, interviews for screening data, MRI acquisitions, and so on. The data obtained is then entered into XNAT by human collaborators or semi automated processes that need human interaction at some point of their workflow, which is prone to introduce errors and inconsistencies in the dataset. Having a sound, error free, dataset is crucial for any data analysis process. Consequently, there is a need for designing a QC strategy that effectively detects and manages this kind of errors. To tackle the QC problem our approach is based on ontology-based data quality management principles. It takes advantage of the logical model defined in the ontologies and expands it with more explicit SPIN rules and constructs.

After transformation, the reasoning step of the ETL pipeline derives data and carries out consistency checks. The reasoner checks the logical restrictions defined in the model to assure data consistency. Simultaneously, the definition of constraints using SPIN rules is also valuable for further and more fine-grained inspections that may be difficult to model using Description Logics alone (Fürber and Hepp, 2010).

The layered approach for the semantic model enables working at different levels of abstraction, which allows to verify raw data from XNAT (e.g., assuring the experiments follow predefined ID patterns) and to control more abstract conceptualizations at the same time.

[16]https://joinup.ec.europa.eu/catalogue/distribution/study-persistent-uris-identification-best-practices-and-recommendations-topic

TABLE 2 | Description of stage categories and simplified criteria definition with Description Logics.

Class	Description	Simplified formal definition
Normal Control (NC)	The subject's MMSE score is over 28, all T-Scores are equal or greater than 35 and does not report subjective cognitive decline	$Normal \equiv StudySubject \sqcap (\exists mmse. \geq 28) \sqcap (\exists TscoreVOSP. \geq 35) \sqcap (\exists TscoreCOWAT. \geq 35) \sqcap (\exists TscoreCERAD\ Recall. \geq 35) \sqcap (\exists TscoreTMTB. \geq 35) \sqcap \neg(\exists reports.SCD)$
Subjective Cognitive Decline (SCD)	The subject's MMSE score is over 28, all T-Scores are equal or greater than 35 and reports subjective cognitive decline	$SCD \equiv StudySubject \sqcap (\exists mmse. \geq 28) \sqcap (\exists TscoreVOSP. \geq 35) \sqcap (\exists TscoreCOWAT. \geq 35) \sqcap (\exists TscoreCERAD\ Recall. \geq 35) \sqcap (\exists TscoreTMTB. \geq 35) \sqcap (\exists reports.SCD)$
Mild Cognitive Impairment (MCI)	The subject's MMSE score is between 23 and 28, having at least one T-Score under 35	$MCI \equiv StudySubject \sqcap (\exists \text{mmse.} > 23) \sqcap (\exists mmse. < 28) \sqcap (\exists \text{Tscore}_{VOSP}. < 35) \sqcup (\exists TscoreCOWAT. < 35) \sqcup (\exists TscoreCERAD\ Recall. < 35) \sqcup (\exists TscoreTMTB. < 35))$
Dementia	The subject's MMSE score is under 23 and has at least one T-Score under 35	$Dementia \equiv StudySubject \sqcap (\exists mmse. \leq 23) \sqcap ((\exists TscoreVOSP. < 35) \sqcup (\exists TscoreCOWAT. < 35) \sqcup (\exists TscoreCERAD\ Recall. < 35) \sqcup (\exists TscoreTMTB. < 35))$

An example of a high-level QC task is finding subjects who meet the exclusion criteria but have not been properly tagged by human supervisors. These errors introduce noise in the data analysis models but are easily overlooked. For this task, ADO defines the class "exclusion criterion," with a set of specific subclasses modeling several exclusion criteria that covered most of the needs of this project. Depending on which of the variables from the subject's medical history experiment are set to true, the subject is related to the specific instance that represents the exclusion. This check is modeled by the SPIN constraint depicted in **Code 1**.

Automatic Staging

A central task within the APGeM project is assessing the subject's stage in cognitive decline for diagnostic purposes and it can be automated based on available screening data stored in XNAT. On the one hand, it is another mean of QC for submitted data, highlighting possible discrepancies between evidence in the screening tests and the final outcome, which may be due to a human error made at data entry or an incorrect diagnosis from the practitioner. On the other hand, it produces useful staging information when the diagnostic interview is missing for any reason. Moreover, the comparison with the manual staging performed by a physician is also noteworthy.

Our approach integrates a simple stage classifier as part of both formal and domain layer. The subject can be staged under 5 different categories, described in **Table 2**. The classifier has been implemented as a set of SPIN rules (**Code 2**) that assess the diagnostic staging by filtering screening data that meets several conditions for different clinical tests.

Reporting and Data Extraction

XNAT provides various means to customize reports and searches to make them accessible through the web interface, such as the advanced use of display files. However, advanced XNAT displaying customization requires good knowledge of the underlying XNAT database structure (for customized SQL views and displays). Also its REST API enables the development of customized scripts. While this method is very powerful for external software development and library design (such as PyXNAT), it requires a fair amount of programming to perform complex queries and data retrieval.

Concept generalization (class subsumption in ontologies) and the graph-based model of RDF provide a powerful and flexible environment for query design. The use of ontologies and SPARQL for "intelligent querying" has been demonstrated many times in the literature (Pathak et al., 2012a,b; Leroux and Lefort, 2015) and is one of the inspirations for the development of our framework. It simplifies the creation of targeted reports and the extraction of subsets of data from different domains for further analysis. For instance, generating CSV files from SELECT clauses or RDF graphs with CONSTRUCT clauses.

Code 3 shows the query employed for tracking subjects that have Diffusion Tensor Imaging (DTI) and are diagnosed with MCI.

DISCUSSION

Comparing the framework to similar approaches is not straightforward, as the benefits are focused in improving development tasks and the assessment may be subjective, dependent on the objectives pursued. We have presented several use cases to illustrate the effectiveness and ease of use of the proposed solution.

The use of ontologies and semantic technologies as a means of data storage, access, and analysis is widely adopted in biomedical projects. However, this type of ventures still comprises a set of challenges. The most time consuming task of them has been the ontology selection, alignment, and mapping. Despite the great availability of different ontologies to the scientific community, many of them overlap in some subsets and/or lack some others, drawing a landscape of competing standards.

The selection of the technologies involved in the transformation, reasoning, and storing of the data is also up to discussion. It is important for the developers to evaluate and find a balance between ease of deployment and performance optimization, which will ultimately depend on

Code 2 | SPIN rule attached to study subject class instances. It constructs new triples to the subject's diagnosis experiment and state Mild Cognitive Impairment at both formal and domain level.

```
PREFIX sio: <http://semanticscience.org/resource/>
PREFIX ado: <http://scai.fraunhofer.de/AlzheimerOntology#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX apgem: <http://www.apgem.org/resource/>

CONSTRUCT{
  # MCI value at formal level
  ?staging sio:SIO_000300 ?inferred.
  # MCI at domain level
  ?staging a ado:Mild_cognitive_Impairment.
}
WHERE {
  # Count T-Scores < 35
 {
   SELECT ?mmsetotal (COUNT(?tscore) AS ?tscorecount)
   WHERE {
     ?this sio:SIO_000062 ?csExperiment.
     ?csExperiment a apgem:apgem_0004 ; sio:SIO_000312 ?csdata.

     ?csdata sio:SIO_000028*/sio:SIO_001277 ?mmse.
     ?mmse rdfs:label "MMSE_Total" ; sio:SIO_000300 ?mmsetota.
     ?csdata sio:SIO_000028*/sio:SIO_001277 ?score.
     ?score rdfs:label ?label ; sio:SIO_000300 ?tscore.
     # The variables must be Tscores
     FILTER (regex(?label, "VOSP_Tscore")
     || regex(?label, "CERAD_Recall_Tscore")
     || regex(?label, "COWAT_Tscore")
     || regex(?label, "TMTB_Tscore")).
     FILTER (?tscore < 35)
   }
  group by ?this ?mmsetotal
}
## data from Medical History experiment
?this sio:SIO_000062 ?mhExperiment.
?mhExperiment a apgem:apgem_0003 ; sio:SIO_000312 ?mhdata.
# Participant informed subjective cognitive decline
?mhdata sio:SIO_000028*/sio:SIO_001277 ?cmhpar.
?cmhpar rdfs:label "P_subcogdec" ; sio:SIO_000300 ?P_subcogdec.
# Informant informed subjective cognitive decline
?mhdata sio:SIO_000028*/sio:SIO_001277 ?cmhinf.
?cmhinf rdfs:label "I_subcogdec" ; sio:SIO_000300 ?I_subcogdec.
# MCI Criteria
FILTER(
  # 23 < MMSE
  23 < ?mmsetotal)
  # Participant or informant cognitive decline
  && (?P_subcogdec != 0 || ?I_subcogdec != 0)
  # One or more t-scores < 35
  && ?tscorecount >= 1)

 # Diagnosis experiment to update
 ?this sio:SIO_000062 ?diagExperiment.
 ?diagExperiment a apgem:apgem_0001 ; sio:SIO_000312 ?diagdata.
 ?diagdata sio:SIO_001277 ?stagnode.
 ?stagnode rdfs:label "stag" ; sio:SIO_000300 ?staging.
 BIND(6 as ?inferred.
}
```

the objectives pursued. Using query rewriting approaches like Ontop (Calvanese et al., 2015) saves development time, but at the expense of performance, which is bound to the complexity of the ontology and mappings. For instance, the rewriting of the queries suffers an exponential blow-up in the worst case (Gottlob et al., 2014). To overcome these problems, the complexity of the ontology needs to be restrained, which would potentially limit the flexibility of the ontological design. Also, the SQL source queries for the mappings need to be as optimal as possible. This task requires good knowledge of both SQL and XNAT database structure. Last but not least, the reasoning capabilities are also limited.

Code 3 | Query for tracking subjects that have Diffusion Tensor Imaging with a specific diagnosis staging.

```
PREFIX sio: <http://semanticscience.org/resource/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX apgem: <http://www.apgem.org/resource/>
PREFIX dicom: <http://purl.org/nidash/dicom#>

SELECT (count (?subject) as ?total)
WHERE {
  ?subject a sio:SIO_000399; sio:SIO_000062 ?session, ?diagnosis.
  ?session a apgem:apgem_0028; sio:SIO_000312 ?sessiondata.
  ?sessiondata sio:SIO_000028*/sio:SIO_001277 ?desc.
  ?desc a dicom:seriesDescription; sio:SIO_000300 ?description.
  # staging information
   ?diagnosis a apgem:apgem_0001; sio:SIO_000312 ?diagdata.
  ?diagdata sio:SIO_001277 ?stag. ?stag rdfs:label "stag"; sio:SIO_000300 ?stagValue.
  # Get only MCI staged subjects with labels starting with D10
  # with MRSessions with ID ending with _1 and have scans with
  # DTI in its series description
  FILTER(?stagValue = 6
    && regex(?subject, "^D10")
    && regex(?description, "DTI")
    && regex(?expLabel, "-1$"))
}
GROUP BY ?subject
```

Regarding the use of the framework, the preliminary applications show promising results. QC is tightly integrated in the data update workflow, enabling the early detection of noisy and inconsistent data, saving a significant amount of time in data inspection. The data exposed in Fuseki's SPARQL endpoint allows data researchers to prepare very specific datasets in less time. As we thought, the preliminary results obtained by the stage classifier have highlighted discrepancies between its output and the actual diagnosis. Further analysis will be necessary to evaluate the source of these disagreements, which may be due to the simple approach of the current staging algorithm, errors in the data or in the diagnostic process. It opens the way for future applications of the framework.

While the implemented semantic environment already fulfills many of our motivations, there is still room for further improvements. One of the immediate enhancements for our framework is the alignment of the formal level with Linked Data Cubes to generate more self-contained datasets for external analysis. This is easily implemented with dedicated SPARQL constructs that translate from one vocabulary to another. The cubes and slices can be optimized to fit specific Machine Learning algorithms, saving intermediate adaptation steps. Another interesting use for the framework would be information retrieval and annotation of free text comments attached to many different experiments. The challenge mainly lies in the multilingual nature of the comments.

Although the development focuses on the XNAT platform, the modeling and techniques applied foster reutilization and are easily generalizable to other of the available archiving solutions for neuroimaging and clinical data. The only requirement would be the adaptation of the transformations and domain specific conceptualizations.

CONCLUSION

We have presented an incremental, modular, and scalable framework that enhances and extends the capabilities of neuroimaging and biobanking systems through the use of semantic technologies. The approach has been exemplified through the XNAT platform in the context of the APGeM project.

The union of schemas, ontologies and services that together enable semantic data access composes the framework. XNAT model, along with XCEDE and complementary schemas, establish the schema level of the framework, providing a suitable means to consume and exchange imaging and clinical research data. The domain level provides the higher level with more abstract concepts, supporting simpler queries and knowledge modeling. The formal level, which works with low-level and raw data/metadata, provides a good toolset for Quality Control and consistency check. Integrating the reasoner in the pipeline allows taking advantage of the formal definitions, generating further assertions about data quality and classifications.

This work shows that following the proposed methodology is possible to enhance non-semantic biomedical research systems with semantic capabilities, improving data management from low-level data to more descriptive logical concepts. The use cases shown confirm the benefits of applying layered semantic descriptions to multi-dimensional datasets, common in the Neuroscience domain, highlighting the convenience of integrating these technologies in current systems updates and future developments.

AUTHOR CONTRIBUTIONS

All authors participated in the conception, design and implementation of the work, and in the drafting and revision of the paper.

FUNDING

This work has been carried out within the following projects: The Pre-clinical genotype-phenotype predictors of Alzheimer's disease and other dementias (APGeM) project. The dementia disease initiation (DDI) Norwegian research council supported translational research project (217780). The work was also supported by a grant from Iceland, Liechtenstein, and Norway through the EEA Financial Mechanism, supported and coordinated by Universidad Complutense de Madrid (Call ABEL-CM-01-2013).

REFERENCES

Albert, M. S., DeKosky, S. T., Dickson, D., Dubois, B., Feldman, H. H., Fox, N. C., et al. (2011). The diagnosis of mild cognitive impairment due to alzheimer's disease: recommendations from the national institute on aging-Alzheimer's association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's Dementia* 7, 270–279. doi: 10.1016/j.jalz.2011.03.008

Ashburner, M., Ball, C. A., and Blake, J. A. (2000). Gene ontology: tool for the unification of biology. *Nat. Genet.* 25, 25–29. doi: 10.1038/75556

Ashish, N., Ambite, J. L., Muslea, M., and Turner, J. A. (2010). Neuroscience data integration through mediation: an (F)BIRN case study. *Front. Neuroinformatics* 4:118. doi: 10.3389/fninf.2010.00118

Bakken, S., Cimino, J. J., Haskell, R., Kukafka, R., Matsumoto, C., Chan, G. K., et al. (2000). Evaluation of the clinical LOINC (Logical Observation Identifiers, Names, and Codes) semantic structure as a terminology model for standardized assessment measures. *J. Am. Med. Inform. Assoc.* 7, 529–538. doi: 10.1136/jamia.2000.0070529

Butt, A. S., and Khan, S. (2014). Scalability and performance evaluation of semantic web databases. *Arab. J. Sci. Eng.* 39, 1805–1823. doi: 10.1007/s13369-013-0753-4

Calvanese, D., Cogrel, B., Komla-ebri, S., Kontchakov, R., and Lanti, D. (2015). Ontop : answering SPARQL queries over relational databases. *Semant. Web* 8, 471–487. doi: 10.3233/SW-160217

Canuel, V., Rance, B., Avillach, P., Degoulet, P., and Burgun, A. (2015). translational research platforms integrating clinical and omics data: a review of publicly available solutions. *Brief. Bioinform.* 16, 280–290. doi: 10.1093/bib/bbu006

Corradi, L., Arnulfo, G., Schenone, A., Porro, I., and Fato, M. (2009). XTENS - an eXTensible environment for neuroscience. *Stud. Health Technol. Inform.* 147, 127–136. doi: 10.3233/978-1-60750-027-8-127

Cote, R. A., and Robboy, S. (1980). Progress in medical information management: systematized nomenclature of medicine (SNOMED). *JAMA* 243, 756–762.

Courtot, M., Gibson, F., Lister, A. L., Malone, J., and Schober, D. (2011). MIREOT : the minimum information to reference an external ontology term. *Appl. Ontol.* 6, 23–33. doi: 10.3233/AO-2011-0087

Duan, L., Xu, M., Chua, T., Tian, Q., and Xu, C. (2003). "A mid-level representation framework for semantic sports video analysis," in *Proceedings of the Eleventh ACM International Conference on Multimedia* (Berkeley, CA), 33–44. doi: 10.1145/957013.957020

Dumontier, M., Baker, C. J., Baran, J., Callahan, A., Chepelev, L., Cruz-Toledo, J., et al. (2014). The Semanticscience Integrated Ontology (SIO) for biomedical research and knowledge discovery. *J. Biomed. Semant.* 5:14. doi: 10.1186/2041-1480-5-14

Fielding, R. T., and Taylor, R. N. (2002). Principled design of the modern web architecture. *ACM Trans. Internet Technol.* 2, 115–150. doi: 10.1145/514183.514185

Fillenbaum, G. G., van Belle, G., Morris, J. C., Mohs, R. C., Mirra, S. S., Davis, P. C., et al. (2008). Consortium to establish a registry for Alzheimer's disease (CERAD): the first twenty years. *Alzheimer's Dement.* 4, 96–109. doi: 10.1016/j.jalz.2007.08.005

Fladby, T., Pålhaugen, L., Selnes, P., Waterloo, K., Bråthen, G., Hessen, E., et al. (2017). Detecting at-risk alzheimer's disease cases. *J. Alzheimer's Dis.* doi: 10.3233/JAD-170231. [Epub ahead of print].

Folstein, M. F., Folstein, S. E., and McHugh, P. R. (1975). 'Mini-Mental State'. A practical method for grading the cognitive state of patients for the clinician. *J. Psychiatr. Res.* 12, 189–198.

Fürber, C., and Hepp, M. (2010). "Using SPARQL and SPIN for data quality management on the semantic web," in *Business Information Systems: 13th International Conference, BIS 2010, Proceedings,* eds W. Abramowicz and R. Tolksdorf (Berlin; Heidelberg: Springer), 35–46. doi: 10.1007/978-3-642-12814-1_4

Gadde, S., Aucoin, N., Grethe, J. S., Keator, D. B., Marcus, D. S., Pieper, S., et al. (2012). XCEDE: an extensible schema for biomedical data. *Neuroinformatics* 10, 19–32. doi: 10.1007/s12021-011-9119-9

Gene Ontology Consortium (2010). The gene ontology in 2010: extensions and refinements. *Nucleic Acids Res.* 38(Suppl. 1), 331–335. doi: 10.1093/nar/gkp1018

Gottlob, G., Kikot, S., Kontchakov, R., Podolskii, V., Schwentick, T., and Zakharyaschev, M. (2014). The price of query rewriting in ontology-based data access. *Artif. Intell.* 213, 42–59. doi: 10.1016/j.artint.2014.04.004

Hastings, J., Frishkoff, G. A., Smith, B., Jensen, M., Poldrack, R. A., Lomax, J., et al. (2014). Interdisciplinary perspectives on the development, integration, and application of cognitive ontologies. *Front. Neuroinform.* 8:62. doi: 10.3389/fninf.2014.00062

Hoehndorf, R., Schofield, P. N., and Gkoutos, G. V. (2015). The role of ontologies in biological and biomedical research: a functional perspective. *Brief. Bioinform.* 16, 1069–1080. doi: 10.1093/bib/bbv011

Hsu, W., Gonzalez, N. R., Chien, A., Villablanca, J. P., Pajukanta, P., Viñuela, F., et al. (2015). An integrated, ontology-driven approach to constructing observational databases for research. *J. Biomed. Inform.* 55, 132–142. doi: 10.1016/j.jbi.2015.03.008

Huff, S. M., Rocha, R. A., and McDonald, C. J. (1998). Development of the Logical Observation Identifier Names and Codes (LOINC) vocabulary. *J. Am. Med. Inform. Assoc.* 5, 276–292. doi: 10.1136/jamia.1998.0050276

Izzo, M. (2016). *Biomedical Research and Integrated Biobanking: An Innovative Paradigm for Heterogeneous Data Management.* Springer Theses.

Jessen, F., Amariglio, R. E., van Boxtel, M., Breteler, M., Ceccaldi, M., Chételat, G., et al. (2014). A conceptual framework for research on subjective cognitive decline in preclinical Alzheimer's disease. *Alzheimer's Dement.* 10, 844–852. doi: 10.1016/j.jalz.2014.01.001

Keator, D. B., Helmer, K., Steffener, J., Turner, J. A., van Erp, T. G., Gadde, S., et al. (2013). Towards structured sharing of raw and derived neuroimaging data across existing resources. *Neuroimage* 82, 647–661. doi: 10.1016/j.neuroimage.2013.05.094

Kilintzis, V., and Beredimas, N. (2014). Evaluation of the performance of open-source RDBMS and triplestores for storing medical data over a web service. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* 2014, 4499–4502. doi: 10.1109/EMBC.2014.6944623

Leroux, H., and Lefort, L. (2015). Semantic enrichment of longitudinal clinical study data using the CDISC standards and the semantic statistics vocabularies. *J. Biomed. Semant.* 6:16. doi: 10.1186/s13326-015-0012-6

Luciano, J. S., Andersson, B., Batchelor, C., Bodenreider, O., Clark, T., Denney, C. K., et al. (2011). The translational medicine ontology and knowledge base: driving personalized medicine by bridging the gap between bench and bedside. *J. Biomed. Semant.* 2(Suppl. 2):S1. doi: 10.1186/2041-1480-2-S2-S1

Malhotra, A., Younesi, E., Gündel, M., Müller, B., Heneka, M. T., and Hofmann-Apitius, M. (2014). ADO: a disease ontology representing the domain knowledge specific to Alzheimer's disease. *Alzheimer's Dement.* 10, 238–246. doi: 10.1016/j.jalz.2013.02.009

Marcus, D. S., Olsen, T. R., Ramaratnam, M., and Buckner, R. L. (2007). The extensible neuroimaging archive toolkit: an informatics platform for managing, exploring, and sharing neuroimaging data. *Neuroinformatics* 5, 11–34. doi: 10.1385/NI:5:1:11

Maumet, C., Auer, T., Bowring, A., Chen, G., Das, S., Flandin, G., et al. (2016). Sharing brain mapping statistical results with the neuroimaging data model. *Sci. Data* 3:160102. doi: 10.1038/sdata.2016.102

McDonald, C. J., Huff, S. M., Suico, J. G., Hill, G., Leavelle, D., Aller, R., et al. (2003). LOINC, a universal standard for identifying laboratory observations: a 5-year update. *Clin. Chem.* 49, 624–633. doi: 10.1373/49.4.624

Mechouche, A., Golbreich, C., and Gibaud, B. (2007). Towards a hybrid system using an ontology enriched by rules for the semantic annotation of brain MRI images. *Web Reason. Rule Syst. LNCS* 4524, 219–228. doi: 10.1007/978-3-540-72982-2_16

Mechouche, A., Morandi, X., Golbreich, C., and Gibaud, B. (2009). A hybrid system using symbolic and numeric knowledge for the semantic annotation of sulco-gyral anatomy in brain MRI images. *IEEE Trans. Med. Imaging* 28, 1165–1178. doi: 10.1109/TMI.2009.2026746

Mejino, J. L., Rubin, D. L., and Brinkley, J. F. (2008). "FMA-RadLex: an application ontology of radiological anatomy derived from the foundational model of anatomy reference ontology," in *AMIA Annual Symposium Proceedings/AMIA Symposium* (Washington, DC), 465–469.

Mitchell, A. J., Bird, V., Rizzo, M., and Meader, N. (2010). Diagnostic validity and added value of the Geriatric Depression Scale for depression in primary care: a meta-analysis of GDS 30 and GDS 15. *J. Affect. Disord.* 125, 10–17. doi: 10.1016/j.jad.2009.08.019

Musen, M. A., and Noy, N. F. (2011). The national center for biomedical ontology. *J. Am. Med. Inform. Assoc.* 19, 190–195. doi: 10.1136/amiajnl-2011-000523

Parsia, B., Matentzoglu, N., Gonçalves, R. S., Glimm, B., and Steigmiller, A. (2017). The OWL reasoner evaluation (ORE) 2015 competition report. *J. Autom. Reason.* doi: 10.1007/s10817-017-9406-8

Pathak, J., Kiefer, R. C., and Chute, C. G. (2012a). "Applying linked data principles to represent patient's electronic health records at mayo clinic: a case report," in *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium* (Miami, FL), 455–464.

Pathak, J., Kiefer, R. C., Bielinski, S. J., and Chute, C. G. (2012b). Applying semantic web technologies for phenome-wide scan using an electronic health record linked biobank. *J. Biomed. Semant.* 3, 1–18. doi: 10.1186/2041-1480-3-10

Rosse, C., and Mejino, J. L. (2003). A reference ontology for biomedical informatics: the foundational model of anatomy. *J. Biomed. Inform.* 36, 478–500. doi: 10.1016/j.jbi.2003.11.007

Samwald, M., Jentzsch, A., Bouton, C., Kallesøe, C. S., Willighagen, E., Hajagos, J., et al. (2011). Linked open drug data for pharmaceutical research and development. *J. Cheminfom.* 3:19. doi: 10.1186/1758-2946-3-19

Scheufele, E., Aronzon, D., Coopersmith, R., McDuffie, M. T., Kapoor, M., Uhrich, C. A., et al. (2014). "tranSMART: an open source knowledge management and high content data analytics platform," in *AMIA Joint Summits on Translational Science Proceedings* (San Francisco, CA: American Medical Informatics Association), 96–101.

Schwartz, Y., Barbot, A., Thyreau, B., Frouin, V., Varoquaux, G., Siram, A., et al. (2012). PyXNAT: XNAT in python. *Front. Neuroinform.* 6:12. doi: 10.3389/fninf.2012.00012

Scott, A., Courtney, W., Wood, D., de la Garza, R., Lane, S., King, M., et al. (2011). COINS: an innovative informatics and neuroimaging tool suite built for large heterogeneous datasets. *Front. Neuroinform.* 5:33. doi: 10.3389/fninf.2011.00033

Shulman, K. I. (2000). Clock-drawing: is it the ideal cognitive screening test? *Int. J. Geriatr. Psychiatry* 15, 548–561. doi: 10.1002/1099-1166(200006)15:6<548::AID-GPS242>3.0.CO;2-U

Sim, I., Tu, S. W., Carini, S., Lehmann, H. P., Pollock B. H., Peleg M., et al. (2014). The Ontology of Clinical Research (OCRe): an informatics foundation for the science of clinical research. *J. Biomed. Inform.* 52, 78–91. doi: 10.1016/j.jbi.2013.11.002

Sirin, E., Parsia, B., Grau, B. C., Kalyanpur, A., and Katz, Y. (2007). Pellet: a practical OWL-DL reasoner. *Web Semant.* 5, 51–53. doi: 10.1016/j.websem.2007.03.004

Sonntag, D. (2008). "Towards dialogue-based interactive semantic mediation in the medical domain," in *The 7th International Semantic Web Conference*. Available online at: http://disi.unitn.it/~p2p/OM-2008/om2008_proceedings.pdf#page=260

Steigmiller, A., Liebig, T., and Glimm, B. (2014). Konclude: System Description. *Web Semant.* 27–28, 78–85. doi: 10.1016/j.websem.2014.06.003

Su, C. J., and Peng, C. W. (2012). Multi-agent ontology-based Web 2.0 platform for medical rehabilitation. *Exp. Syst. Appl.* 39, 10311–10323. doi: 10.1016/j.eswa.2011.09.089

Suárez-Figueroa, M. C., Gómez-Pérez, A., and Fernández-López, M. (2012). "The NeOn methodology for ontology engineering," in *Ontology Engineering in a Networked World,* eds M. C. Suárez-Figueroa, A. Gómez-Pérez, E. Motta, and A. Gangemi (Berlin; Heidelberg: Springer), 9–34. doi: 10.1007/978-3-642-24794-1_2

Whetzel, P. L., Noy, N. F., Shah, N. H., Alexander, P. R., Nyulas, C., Tudorache, T., et al. (2011). BioPortal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. *Nucleic Acids Res.* 39, W541–W545. doi: 10.1093/nar/gkr469

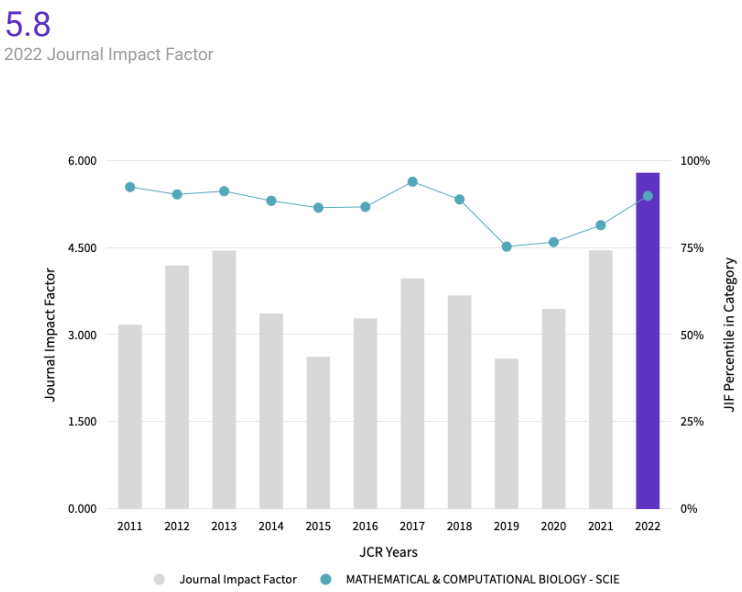
Zhang, S., Bodenreider, O., Li, A., Mejino, J., Agoncillo, A., Brinkley, J., et al. (2003). Law and order: assessing and enforcing compliance with ontological modeling principles in the foundational model of anatomy. *Comput. Biol. Med.* 36, 674–693. doi: 10.1016/j.compbiomed.2005.04.007

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Chapter 4

An overview of graph databases and their applications in the biomedical domain

Title	*An overview of graph databases and their applications in the biomedical domain*
Journal	Database-The Journal of Biological Databases and Curation
Authors	Santiago Timón, Mariano Rincón, and Rafael Martínez Tomás
Published	30 April 2021
Impact Factor	4.462
JCR Quartile	Q1
DOI	10.1093/database/baab026



DATABASE
The Journal of Biological Databases and Curation

Database, 2021, 1–22
doi:10.1093/database/baab026
Review

Review

An overview of graph databases and their applications in the biomedical domain

Santiago Timón-Reina*, Mariano Rincón and Rafael Martínez-Tomás

Departamento de Inteligencia Artificial, Universidad Nacional de Educación a Distancia (UNED), C/Juan del Rosal, 16 Ciudad Universitaria, Madrid 28040, Spain

*Corresponding author: Tel: +34 91-398-7209; Email: santiagotimon@dia.uned.es

Citation details: Timón-Reina, S., Rincón, M., Martínez-Tomás, R. *et al.* An overview of graph databases and their applications in the biomedical domain. *Database* (2021) Vol. 2021: article ID baab026; doi:10.1093/database/baab026

Received 28 November 2020; Revised 24 March 2021; Accepted 30 April 2021

Abstract

Over the past couple of decades, the explosion of densely interconnected data has stimulated the research, development and adoption of graph database technologies. From early graph models to more recent native graph databases, the landscape of implementations has evolved to cover enterprise-ready requirements. Because of the interconnected nature of its data, the biomedical domain has been one of the early adopters of graph databases, enabling more natural representation models and better data integration workflows, exploration and analysis facilities. In this work, we survey the literature to explore the evolution, performance and how the most recent graph database solutions are applied in the biomedical domain, compiling a great variety of use cases. With this evidence, we conclude that the available graph database management systems are fit to support data-intensive, integrative applications, targeted at both basic research and exploratory tasks closer to the clinic.

Introduction

Nowadays, the generation, consumption and, more importantly, analysis of highly interconnected data have become ubiquitous. In this situation, where the relationships among data grow both in quantity and in significance, graph models become an appealing solution, as graphs are mathematical entities in which objects are connected. Formally, a graph $G(V, E)$ is composed of an ordered pair of two disjoint sets: vertices V (also referred to as nodes) and edges (or links) E (1). The graph abstraction directly translates concepts and instances into nodes and their relationships into edges, making it intuitive for data modeling. However, strong graph data is not straightforward in conventional Database Management Systems (DBMSs), and the physical implementation of a given data model and how the relations are treated ultimately depend on the database type.

For example, the basis of Relational Database Management Systems (RDBMSs) are tables (relations) (2–4), where each row represents a single data element of an entity and a single column usually defines a particular data attribute. The standard mechanism to create relationships between entities is by defining unique IDs (primary keys) that can be copied into referencing tables (foreign keys). To exploit these references and include different tables in a database query, the Structured Query Language (SQL) (5) provides

the JOIN clause. The relational paradigm is very appropriate for well-defined data structures that are unlikely to change and translate naturally to tables, and the relations among its entities are not numerous and not as relevant as the entities' attributes. Hence, given its maturity and technological development, RDBMSs are widely used for data storage, with countless examples experienced in everyday life, like user data, inventory tracking, blog posts and many more. However, when most relationships are many-to-many, prevalent in densely connected data, querying the database requires multiple expensive JOIN operations, impacting the performance (6).

Although graphs can be modeled with tables representing vertices and edges, complex queries or graph algorithms (like path traversals) are challenging to optimize without implementing complementary structures, such as adjacency lists (7). These modeling and performance limitations have increased the interest in Graph Database Management Systems (GDBMSs). GDBMSs, in contrast to regular DBMSs, allow working directly with a graph model, avoiding sophisticated engineering to represent relationships efficiently, and provide straightforward ways to store, access and operate graph data, especially for traversing paths and matching subgraphs. Furthermore, the schema-less or schema-optional approach that most GDBMSs follow grants a high degree of flexibility, allowing applications to adapt and evolve quickly and introduce abstraction, specialization of entities and relations among them more easily.

Graph models are present in multiple formal representations and become very powerful when the problem model exhibits varied relations among the entities or concepts. Consequently, the trend in graph databases has permeated into many disparate domains, and we can find applications in Energy Management Systems (EMS) (8), Power Grid Modeling (9) and even less technologically driven fields like Digital Humanities (10). The biomedical domain is a complex area that is inevitably studied in many different sub-domains that are inherently related and connected. For instance, the study of human metabolism requires identifying hundreds of concepts (e.g. metabolites, proteins, complexes and metabolic reaction names) and the relations among them (e.g. consumption, production and catalysis), and graph models provide a valuable framework in this situation. Moreover, the amount of data produced in the 'omics' era results in large graphs that become difficult to manage without a database optimized for the task.

We can illustrate the differences between the relational and graph-based paradigms depicted in Figures 1 and 2, a stripped-down biological model describing subject diagnoses and their related phenotype–genotype and pathway implications. For most GDBMSs, the physical design resulting from the logical model described in Figure 1 would be almost equivalent. However, in the case of RDBMSs, the implementation from the logical to the final physical design requires dealing with the many-to-many cardinality that most of the model's relations will have. A typical normalized relational design, at least to the Third Normal Form (3NF) (11), prevents data redundancy by introducing intermediate tables for each relationship between two entities, as shown in Figure 2. For searching heavily connected entities, like genes, this layout would require referencing (joining and sub-querying) several tables multiple times, potentially with various filters, ultimately eroding the query's performance. Also, complicated queries may end up being rather cumbersome. Thus, designing a relational model for highly interconnected data poses an engineering challenge, especially when the model requires fine-grained semantics, which involves a trade-off between implementing specialized relations (more tables) or limiting the expressiveness at the expense of semantics.

GDBMSs treat relationships as first-class objects, improving the data model's semantics and easing the adoption of knowledge models and ontologies, which are computer science constructs that provide well-defined vocabularies that allow the precise and machine-readable description of knowledge about a particular domain (12). The biomedical domain has driven and benefited from advances in Knowledge Representation (KR) and storage, being one of the early adopters of ontological research. As a result, there exists a significant number of formal biomedical ontologies (13) that capture and model knowledge from disparate sub-fields, giving rise to initiatives like the Open Biological and Biomedical Ontology (OBO) Foundry (14) and the National Center for Biomedical Ontology (15) to promote harmonization and interoperability. These controlled vocabularies and ontologies support the research in several ways, mainly in data annotation (16–19) and biomedical text mining (20, 21).

In this paper, we survey the adoption of GDBMSs in the biomedical domain to present a summary review from an 'application perspective' with categorization and description of biomedical applications employing GDBMSs as storage systems. The applications presented are selected from a broad literature search complying with the following characteristics: (i) are biomedical applications using GDBMSs, (ii) are well documented with papers and websites (iii) have been peer-reviewed. Our coverage of biological graph-powered systems is by no means exhaustive, focusing on recent developments that are high quality, publicly available and expected to be of interest to experts and developers in the community. It is worth noting that, given the overlapping nature of biomedical knowledge,

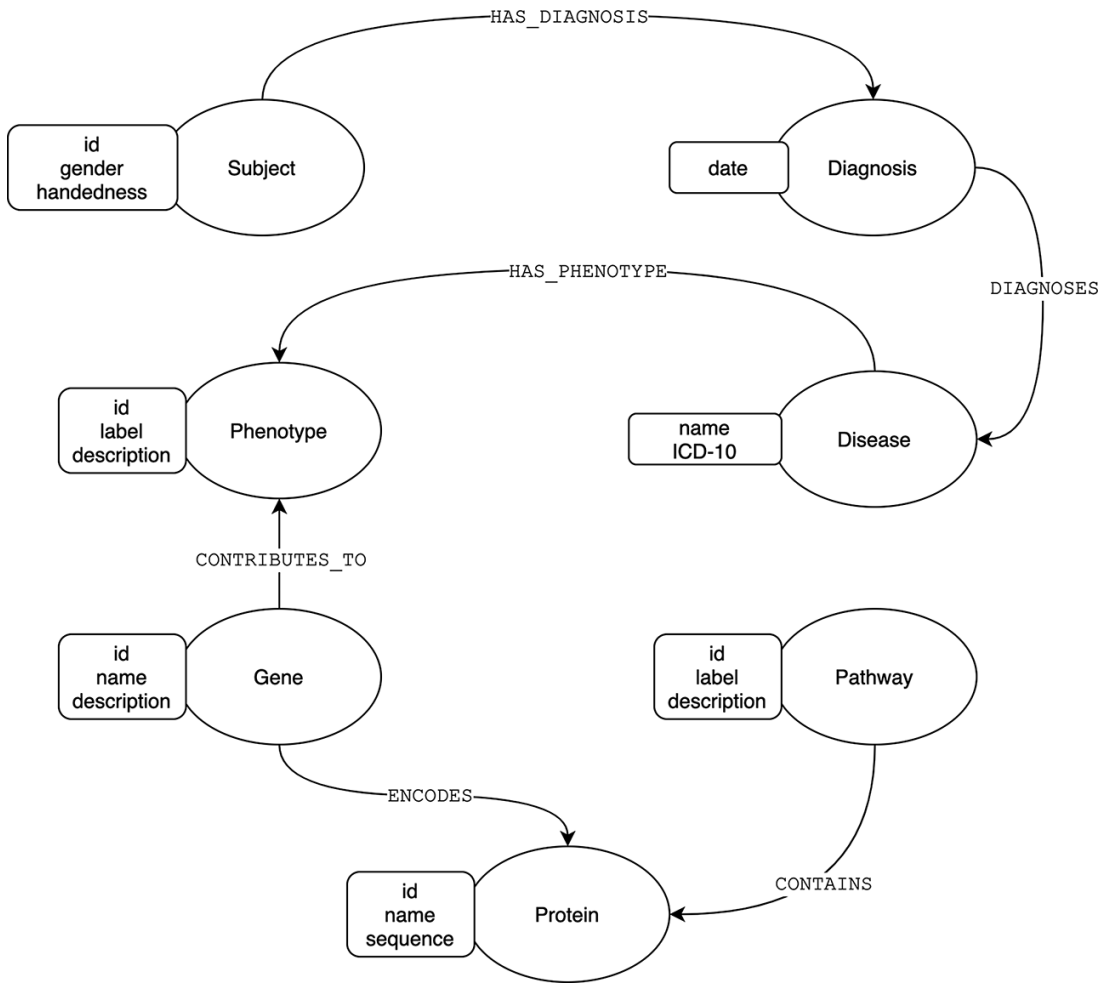


Figure 1. Graph model of diagnoses and its related phenotype–genotype and pathway implications.

some systems can be classified into more than one category. First, we provide a technological background by exploring the different database models and designs and examining the performance through benchmark studies from the literature. Afterward, we highlight the use of GDBMSs within different applications in a wide variety of biomedical contexts, describing the implications and impact of graph technology in these settings. Finally, we discuss the current state, limitations and possible future lines.

Background

Graph database models and design

Graph database models may be defined as those in which the data structures are modeled as a directed, possibly labeled, graph, or its generalizations. The data manipulation is done using graph-oriented operations and type constructors, and appropriate integrity constraints can be defined over the graph structure (22). Over the past decade, graph database implementations have grown from prototypical, application-driven approaches to fully developed products, providing external interfaces, database languages, query optimizers, storage and transaction engines, and management features. This evolution has been actively reviewed (23–28), showing how deficiencies such as the lack of integrity constraints, partition and scalability limitations, or the need for standard graph database languages have been addressed throughout the version history. Besta *et al.* describe the contemporary technological landscape of graph database solutions through a taxonomy of six key design aspects: type of backend technology, data modeling approach, internal data organization, data distribution, query execution and type of transactions (29).

As far as backend technology is concerned, we can see that, at present, most graph database systems are built upon existing storage designs from both relational and NoSQL (30) paradigms, such as key-value, document, wide-column, tuple and object-oriented stores. Key-value stores allocate items as (key, value) pairs, usually in

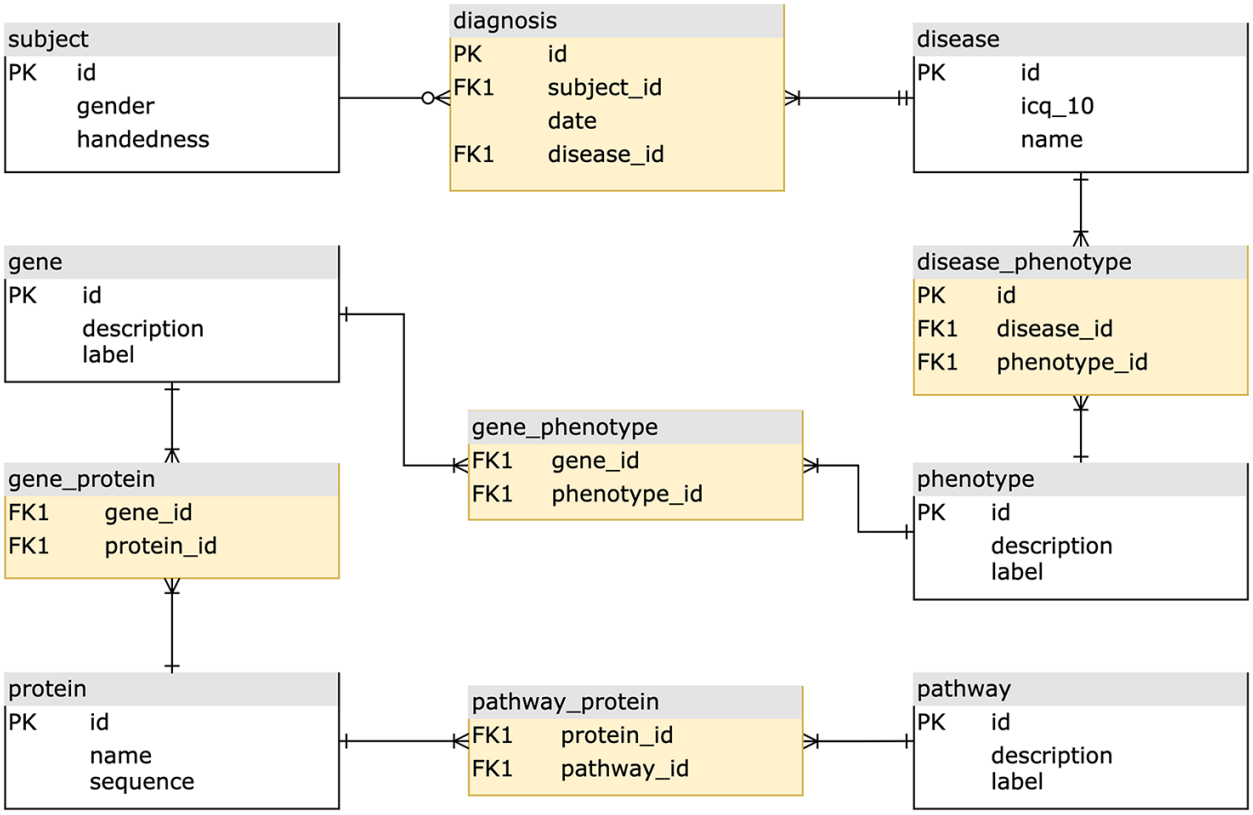


Figure 2. The equivalent normalized relational physical design, with entity tables (white) to store attributes, and join tables (yellow) to implement the relationships.

standalone hash tables. Document stores extend key value so that the values are 'documents', encoded in standard semi-structured formats such as XML, JSON or BSON (Binary JSON). Wide-column stores represent data through a tabular format of rows with a fixed number of column families (an arbitrary number of columns that are logically related to each other and usually accessed together). Triple stores [also known as Resource Description Framework (RDF) databases] work with the notion of triples (subject–object–predicate), and tuple stores generalize these systems to collect tuples of arbitrary size. Object-oriented stores store data as true objects, identified by object IDs (OIDs) and following a class hierarchy. Using existing engines delivers the advantage of mature and well-tested technology but at the expense of obtaining non-optimized graph data representations and queries. In contrast, native graph databases like TigerGraph (31) and Neo4j are specifically built to maintain and process graphs. Table 1 provides a list of different GDBMSs, which many of the reviewed applications use, with their internal database engines.

Regarding data modeling, Labeled Property Graphs (LPG) and RDF are the most common graph models found in graph database systems (32–34). LPG augments the simple graph model to allow defining labels for nodes and edges, as well as an arbitrary number of properties (also called attributes) for both. RDF, a World Wide Web Consortium (W3C) standard, was conceived as a collection of specifications for representing information to allow easy data exchange between different data formats, and graphs arise from the collection of triples in the form of subject, predicate and object (s, p, o). The RDF format is widely used in biomedical setups, due mainly to the fact that RDF is a serialization and data instantiation format for OWL-based bio-ontologies, and new systems using native graph databases rely on transformations between models to fully exploit their features.

Likewise, systems need to define data structures to represent graphs in the storage layer. The most common representation formats are the adjacency matrix (AM), the adjacency list (AL) and the edge list (EL). Figure 3 shows a graphical representation of these formats. The AM is a square matrix where its cells indicate whether vertex pairs are adjacent (connected) or not. In the AL format, each vertex has an associated adjacency list containing the IDs of all adjacent vertices. The difference with EL is that AL explicitly stores edges with its source and destination vertex. The AL format is efficient on traversal operations, and

Table 1. Summary of available implementations by core database engine

Product	Link	Database engine
WhiteDB	http://whitedb.org	Tuple store
GraphDB	https://www.ontotext.com/products/graphdb	Tuple store
OrientDB	https://www.orientdb.org	Document store
ArangoDB	https://www.arangodb.com	Document store
Azure Cosmos DB	https://azure.microsoft.com/es-es/services/cosmos-db	Document store
FaunaDB	https://fauna.com	Document store
RedisGraph	https://oss.redislabs.com/redisgraph	Key-value store
Dgraph	https://dgraph.io	Key-value store
HyperGraphDB	http://www.hypergraphdb.org	Key-value store
MS Graph Engine	https://www.graphengine.io	Key-value store
Titan	https://titan.thinkaurelius.com	Wide-column store
JanusGraph	https://janusgraph.org	Wide-column store
DSE Graph	https://www.datastax.com/products/datastax-graph	Wide-column store
InfiniteGraph	https://www.objectivity.com/products/infinitegraph	Object-oriented store
ThingSpan	https://www.objectivity.com/products/thingspan	Object-oriented store
VelocityDB	https://velocitydb.com	Object-oriented store
Oracle Spatial and Graph	https://www.oracle.com/technetwork/database-options/spatialandgraph/overview/spatialandgraph-1707409.html	RDBMS
Sparksee/DEX	http://www.sparsity-technologies.com	Native graph database
TigerGraph	https://www.tigergraph	Native graph database
GraphBase	https://graphbase	Native graph database
Memgraph	https://memgraph.co	Native graph database
Neo4j	https://neo4j.com	Native graph database

many graph databases use it. Other features, such as index support, are also relevant for the overall performance.

Data distribution may be achieved through data replication or sharding. With replication, each instance maintains a copy of the dataset, while sharding fragments the data across instances. Distribution becomes essential when dealing with large amounts of data, and query execution is directly linked to it. Multi-server query execution can be enabled in several ways. The concurrent execution allows the execution of different queries at the same time, providing higher throughput. With parallelization, a single query can be executed across servers to obtain lower latencies. Because managing large amounts of data can compromise the system's performance or availability, these features can become essential for projects in this situation.

Finally, GDBMSs can be evaluated by the support of transactions. Specifically, Atomicity, Consistency, Isolation, Durability (ACID); Online Transaction Processing (OLTP); and Online Analytics Processing (OLAP) support. OLTP systems focus on smaller transactional queries, while OLAP systems execute more expensive analytic queries that span whole graphs.

The literature reveals that the field is evolving rapidly and many referenced databases have either already been discontinued or greatly improved at the time of writing.

Performance and benchmarking

Because of their innate capabilities in dealing with highly interconnected data, graph databases have been attracting attention in the past years. As different technological implementations of graph database engine have emerged, so has the need for accurate, quantitative performance comparisons between them by using standardized queries and workloads. Furthermore, the differences in relational and graph-based paradigms also raised questions about how they would behave in different contexts. Table 2 summarizes the surveyed benchmark studies.

Within standard benchmarks, the Linked Data Benchmark Council (LDBC) (35) is one of the most consistent works in this topic, and its workloads have been employed and adapted in many benchmarking studies. The library currently includes three kinds of workloads: interactive, business intelligence and graph analytics. Interactive workloads focus on general graph database operations, executing read-only (short and complex) and transactional update queries. Business Intelligence workloads are designed to stress different performance aspects, employing read-only aggregation operations over significant volumes of data that span large parts of the graph. The last workload, 'graphalytics' (36), proposes six graph algorithms to enable the objective comparison of graph analysis platforms: Breadth-First Search (37), PageRank (38),

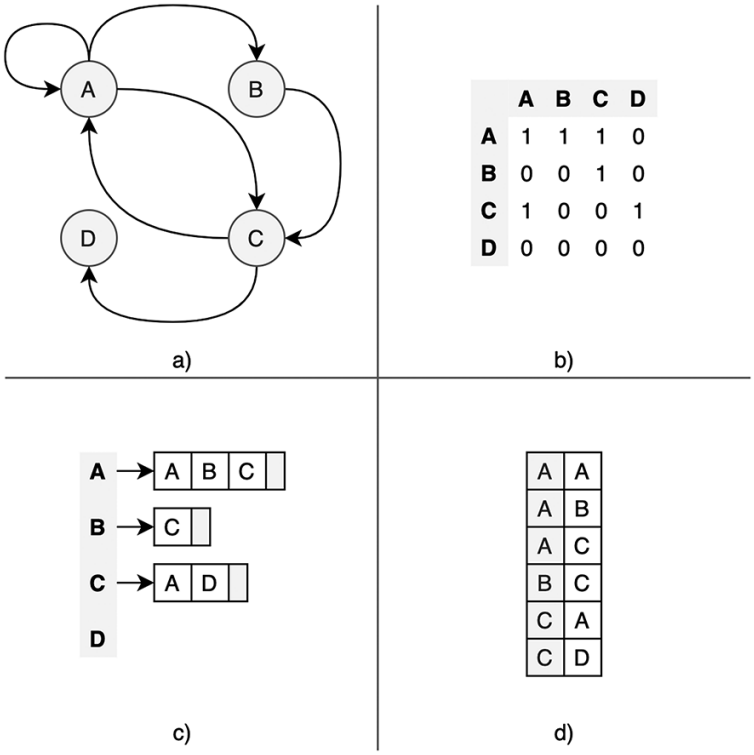


Figure 3. Graphic description of the most common graph representation formats. (a) Original directed graph; (b) adjacency matrix; (c) adjacency list; and (d) edge list.

weakly connected components (39), community detection using label propagation (40), deriving the local clustering coefficient (41), and computing single-source shortest paths.

GDBMSs have been assessed in studies from different contexts, like data provenance (42), biomedical settings (43–46) and social networks (47–52). Most of the social network benchmarks use or adapt the LDBC's Social Network Benchmark (SNB) (53). In parallel with technological surveys, these studies show how GDBMS technology has matured and grown into a competitive and heterogeneous environment, with its weaknesses and strengths.

The number of edges involved in a query has a big impact on performance (44, 46). Likewise, subgraph-matching queries are more challenging to handle in large datasets, in contrast to traversal queries employed in some of the works. Lastly, GDBMSs are, in general, less optimized for aggregate operations (25, 51, 52, 54). In contrast, all the studies acknowledge that schema-less provides a high degree of flexibility to accommodate new nodes or relations, avoiding the need to restructure the schema. GDBMSs are more efficient traversing large graph instances, with lower computational cost than RDBMSs (42, 43, 45, 47, 52, 55, 56), because the search space is reduced to directly connected nodes, avoiding scanning the entire graph to find the nodes that meet the search criteria. Furthermore, graph algorithms (e.g. pathfinding, community detection, centrality or similarity) are more natural to implement and even available out of the box, like the case of Neo4j's Graph Data Science Library (https://neo4j.com/graph-data-science-library/) or TigerGraph's (31) GSQL Graph Algorithm Library (https://docs-beta.tigergraph.com/tigergraph-platform-overview/graph-algorithm-library).

To compare different paradigms, benchmarking implementations require an extra effort to address peculiarities. In the case of RDBMSs vs. GDBMSs (52), Cheng *et al*. propose a unified benchmark that extends the TPC-H (http://www.tpc.org/tpc_documents_current_versions/pdf/tpc-h_v2.18.0.pdf) standard RDBMS benchmark and LDBC using transformation mechanisms between relational and graph data, making it possible to evaluate different systems on the same datasets, query workloads and metrics. The query workloads consist of three main categories. Firstly, atomic relational queries (Projection, Aggregation, Join and Order by) aim to evaluate the performance of primitive relational operations implemented in GDBMSs. Secondly, TPC-H query workloads evaluate the performance of GDMBSs on operations that

Table 2. Relevant benchmarking studies

Reference	Benchmark/ methodology	GDBMS	RDBMS/NoSQL	Description
(42)	Own implementation	Neo4j	MySQL	Technology comparison about recording and querying data provenance information. Executes objective benchmarks to measure query response time and disk space usage. Also, it provides subjective comparisons based on system documentation and usage experience. Concludes that Neo4j outperforms on structural queries, but it is premature to use a graph database in production environments for data provenance
(57)	HPC Scalable Graph Analysis Benchmark	Neo4j Jena Hypergraph DBDEX		Evaluates the performance of selected systems with the HPC benchmark. This benchmark employs R-MAT (58) for graph generation and measures the execution time over different kernels: data loading, scan edges, 2-hops subgraph building and Traversed Edges Per Second. All four platforms perform well on small graphs. Only DEX and Neot4j were able to load the largest graphs. DEX showed the best performance
(47)	Own implementation of a small social network–like problem	Neo4j	MySQL	Small comparative analysis with social network queries. In this study, Neo4j outperforms MySQL in all queries
(59)	GDB, an extensible tool to compare Blueprints-compliant graph databases	Neo4j DEX Titan OrientDB		A Tinkerpop-based distributed benchmarking framework to compare Blueprints-compliant graph databases. The benchmark measures traversal, load and intensive workloads. The results show that all databases perform similarly on read-only operations, while Titan and DEX stood out on read–write workloads, and Neo4j did on traversal workloads. Code available at https://github.com/Jsalim/GraphDB-Benchmark
(43)	Bioinformatics graph processing problems	Neo4j	PostgreSQL	A query benchmark that evaluates Neo4j against PostgreSQL in typical bioinformatics graph processing problems. The study employed the human interaction network from STRING v9.05 (60) and measured the response time for finding immediate neighbors and their interactions, finding the best scoring path between two proteins and finding the shortest path between them. Neo4j outperformed PostgreSQL, showing speedups of 36× (immediate neighbors), 981× (best scoring path) and 2441× (shortest path)
(44)	Graph-based extension to Conditional random field Protein–Protein Interface identification	Neo4j	Microsoft SQL Server	A case study on how Neo4j can be applied to the bioinformatics problem of protein–protein interface identification
(45)	Biomedical graph traversal operations	Neo4j	MySQL	Compares the performance by employing biological network information from 21 different datasets and ontology resources. The benchmark measured the query response time of retrieving all data that traverse the relationships among genes, drugs and diseases that increased the expression of the BRCA1 gene. The results report that Neo4j outperformed MySQL in all cases and highlights the importance of system tuning to obtain better performance

(continued)

Table 2. (Continued)

Reference	Benchmark/ methodology	GDBMS	RDBMS/NoSQL	Description
(49)	LDBC SNB	Neo4j		Analyzed the fundamental points of graph databases and employed the LDBC-SNB to evaluate the performance of Neo4j. Without much detail, the work concludes that Neo4j shows acceptable behavior when dealing with different sizes of graph databases.
(55)	Own implementation of comparative measures over a medical care setup	Neo4j	Oracle	Compares the performance of Oracle 11g and Neo4j over a hospital health-care system use case with a set of predefined queries with different join/subquery requirements. Neo4j outperformed Oracle in 4/5 of the queries.
(50)	Extension of the LDBC SNB IW to simulate real-time transactional workloads	TitanDB Neo4j Virtuoso	PostgreSQL	An improved graph database benchmarking architecture for real-time transaction processing built upon LDBC-SNB and Apache Kafka (https://kafka.apache.org/). Provides LDBC-SNB reference implementations for Gremlin (61), SQL and Cypher. The experiment employed two synthetic datasets with scale factors of 3 and 10, to execute read-only graph queries (point lookups, one-hop traversals, two-hop traversals and single-pair shortest path) and simulate a real-time Interactive Workload. Their results showed that Neo4j achieved higher throughput than TitanDB and that PostgreSQL provided the best overall performance followed by Virtuoso (https://virtuoso.openlinksw.com/) (SQL mode). Concludes that RDBMSs with a native SQL interface provides the best performance under real-time streaming scenarios. Gremlin Server incurs significant overhead
(56)	Follow-up of Khan 2017 with database tuning	Neo4j	Oracle	Follow-up work of Khan 2017 where they improve the performance of Oracle 11g database about 35% by creating separate tablespaces for each schema and table, and five more query workloads. Still, despite the physical tablespace tuning technique of Oracle 11g, Neo4j outperforms it in all proposed scenarios
(52)	Domain-agnostic workloads. Two-way adaptations to compare graph databases with other implementations. TPC-H and LDBC	Neo4j ArangoDB	MySQL Microsoft SQL Server Oracle PosgreSQL RocksDB HBase Cassandra	Comparative evaluation between RDBMSs and GDBMSs under a unified benchmark that extends the TPC-H standard RDBMS benchmark and LDBC. The query workload consists of three main categories: atomic relational queries (projection, aggregation, join and order by), TPC-H query workloads, and five graph algorithms from LDBC. The metrics measured the average query processing time, memory usage (peak) ratio and CPU usage (peak) ratio of five query runs. This benchmark concluded that RDBMSs outperform GDMBSs by a substantial margin under the workloads that mainly consist of group-by, sort and aggregation operations. On the other hand, GDMBSs are superior in the execution of those workloads that mainly consist of multi-table join, pattern matching and path identification

(continued)

Table 2. (Continued)

Reference	Benchmark/ methodology	GDBMS	RDBMS/NoSQL	Description
(51)	Complete LDBC-SNB implementation for Neo4j and TigerGraph	Neo4j TigerGraph		A complete implementation of the LDBC-SNB benchmark in Neo4j and TigerGraph native GDBMSs. The experimental setup consisted of four scale factors that ranged from 1 GB to 1 TB deployed on three computing architectures. The results can be fairly summarized in three key points: TigerGraph stores graph data considerably more compactly than Neo4j, Neo4j is faster at ingesting raw data than TigerGraph, and lastly, Neo4j is faster than TigerGraph only in 13 of the 368 configurations. Concludes that TigerGraph is superior to Neo4j on the LDBC-SNB benchmark
(46)	Own benchmark implementation with biomedical data. Data loading, path traversal and aggregation tests	Neo4j		Evaluates the aptness of the database system in terms of analysis and visualization of a GRN by measuring three test cases (bulk data insertion, path queries and aggregation queries) with a small and large dataset. The results showed that Neo4j performed well in most of the tests; after warming up the cache, the performance improved drastically, reducing query time by about 64% for both dataset sizes. In the same vein (44), the queries that involved more edge operators performed worst
(62)	TigerGraph's benchmark	RedisGraph TigerGraph Neo4j Neptune JanusGraph ArangoDB		The study executes TigerGraph's benchmark to evaluate RedisGraph against leading graph databases. Using graph data from Twitter and Graph500 generator, the benchmark measures the query response time for k-hop neighborhood count, $k = 1,2,3$ and 6. RedisGraph outperforms all competitors, and the study highlights additional opportunities for enhancement: aggregations, enhanced GraphBLAS (http://graphblas.org), Cypher clauses/functionality to support more diverse queries

legacy RDBMSs perform well. And lastly, graph query workloads composed of five graph algorithms in the LDBC Benchmark aimed to evaluate the performance of RDBMSs under the situations GDBMSs are supposed to be efficient.

Nevertheless, on the dichotomy between RDBMSs and GDBMSs, we find how late benchmarks show equivalent or even better performance of the former in different settings, questioning whether it is appropriate to favor GDBMSs over RDBMSs without a proper evaluation of the context. We can find one example in real-life high-throughput scenarios, like those with critical concurrent access (59) or streaming transactional workloads (50), where GDBMSs are less prevalent. In these settings, RDBMSs can deliver competitive performance for OLTP-like online social networking applications, especially in single-node setups. Moreover, the implementation and optimization of graph analytics in RDBMSs are growing areas of research (63–66).

The physical data persistence strategy impacts the overall performance in both paradigms. For example (50), Pacaci *et al*. show how similar SQL queries over the same database schema drive different performance in PostgreSQL and Virtuoso (SQL). The difference is attributable to the fact that Virtuoso employs columnar storage, which is known to suffer under transactional workloads with frequent updates, while PostgreSQL implements row-oriented storage. In the case of GDBMSs, adjacency lists are common in native graph storage, as they enable index-free adjacency access and provide apparent advantages for read operations. However, other storage approaches offer better performance regarding write operations, as is the case of key-value storage engines implementing the LSM-tree (67) index. Moreover, tuning procedures are

Table 3. A short list of useful graph-oriented open-source tools and utilities

Tool	Reference	Code	Description
STON	(74)	https://sourceforge.net/projects/ston	Java-based framework for transforming SBGM models to graphs
Pheno4J	(75)	https://github.com/phenopolis/pheno4j	Java library to load patient genetic variants and phenotype data into Neo4j
Recon2Neo4j	(76)	https://github.com/ibalaur/Recon2Neo4j	Java library that allows loading SBGM models into Neo4j and parsing for translating the Neo4j JSON networks into SBML and SIF formats
ANIMA	(77)	https://github.com/adeffur/ANIMA	R framework for producing multiscale association networks, loaded in Neo4j
SciGraph		https://github.com/SciGraph/SciGraph	Neo4j backed ontology store
Dipper		https://github.com/monarch-initiative/dipper	Python package to generate RDF triples from common scientific resources. Includes mappings and parsers for many sources from different domains
NSMNTX		https://github.com/neo4j-labs/neosemantics	Neo4j plugin that enables the use of RDF in Neo4j
Tarql		https://github.com/tarql/tarql	Java and Apache ARQ based command-line tool for converting CSV files to RDF using SPARQL 1.1 syntax
RDF2Neo	(78)	https://github.com/Rothamsted/rdf2neo	Java-based project providing configurable components to convert RDF data into Cypher commands that can populate a Neo4j graph database

of utter importance to achieve the best possible performance regardless of the system, like optimizing indexing or tablespaces, as some studies report.

Graph database applications in the biomedical domain

Biomedical research produces large amounts of densely interconnected data belonging to many different domains, and storing such data has always presented a technological challenge. Storing graphs using traditional relational databases presents several drawbacks. Relational databases rely on fixed schemas and usually require redesigns when introducing new data structures, affecting flexibility, efficiency and scalability. More generic data models would require many intermediate tables to represent many-to-many relationships, degrading the overall performance because of the need for multiple join operations to traverse interconnected networks. As graph databases matured, they started to gain more attention in the bioinformatics community, given the ubiquity of graphs in this domain. Consequently, many tools emerged to interoperate between formats and paradigms. Table 3 brings together some of the most relevant ones.

The evolution of Knowledge Representation technologies and, more specifically, ontology languages like OWL, enables more complex and interconnected models. Although many of these tools do not necessarily use an explicit graph model, it is commonly implicit in the semantics, opening the door to exploit graph features. One remarkable example of this approach is the Open Biomedical Ontologies (14), which many of the works we are about to describe employ as foundational models. Table 4 summarizes publicly available graph-powered systems.

Applications in systems biology

Intrinsically, systems biology models encode networks of entities and biological processes, such as reactions. As advances in molecular biology produce more extensive and complex networks, the computational demand for analyzing those increases drastically. Consequently, the use of in-house software and desktop solutions started to become a bottleneck. GDBMSs allow decoupling a significant part of the computational needs to dedicated server machines, providing improved tuning of resources for optimal query and algorithm execution performance. One good example is cyNeo4j (68), a Cytoscape (69, 70) app to link this popular network analysis desktop program to a server environment using Neo4j. It enables the user to upload network data and run algorithms both

Table 4. Publicly available graph-powered Biomedical data systems

Platform	Reference	Domain/scope	Implementation	Interfaces	Database
Arena-Idb	(79)	Genetics			Hybrid MySQL-Neo4j
cyNeo4j	(68)		Cytoscape App	Cytoscape GUI	
HRGRN	(80)	Genetics Metabolomics	Web platform	Web	
miTALOS	(81)	Pathway analysis	Web platform	Web	
Biochem4j	(82)	Biochemistry	Exposed database	Neo4j browser REST	Neo4j
Recon2Neo4j	(83)	Metabolomics Proteomics	Exposed database	Neo4j browser	
GeNNet	(84)	Transcriptome analysis	Local web platform	Web Database interface	
Monarch Initiative	(85)	Phenotype–Genotype analysis		Web Data endpoint Ontology endpoint	
Reactome	(86, 87)	Molecular biology Pathway analysis	Web platform	Web Cypher interface REST	
Spfy	(88)	Bacterial WGS	Web platform	Web	Hybrid Blazegraph-MongoDB
GREG	(89)	Genetics	Web platform	Web Cypher interface	Neo4j

locally and on the Neo4j server, creating an interactive workflow that uses the computational strength of the Neo4j server without interrupting the typical workflow in Cytoscape.

Standard formats of the domain, like Systems Biology Markup Language (SBML) (71) or CellML (72), enable modeling biological systems in terms of functional, behavioral or structural aspects, including meta-data and semantic annotations to relate model entities to external resources describing the underlying biology. These meta-data are of great importance to facilitate model reuse and reproducibility, but this introduces heterogeneity, which complicates the design in fixed-schema database systems (73). Henkel, Wolkenhauer and Waltemath employed Neo4j to store SBML and CellML models, including ontology terms and relations from the semantic annotations that these formats support, effectively combining computational models, semantic annotations and simulation experiments. The approach integrated widely adopted bio-ontologies, adding all classes and relations as nodes and edges but leaving out cross-references between concepts of different ontologies. This integration allows querying the information hidden in the semantic annotations of in-model representations and simulation descriptions. Furthermore, it allows defining flexible connections between the data domains, incorporating links between annotations, whole models and model entities.

The Systems Biology Graphical Notation (SBGN) (90) is another standard for visual representation of biological networks. It is composed of three orthogonal languages for representing different views of biological systems: Process Descriptions (PDs), Entity Relationships (ERs) and Activity Flows (AFs). SBGN-to-Neo4j (STON) (74) is a Java framework to transform SBGN markup language files into a Neo4j graph representation, focused only on the PD and AF sub-languages. The authors report that the persistent graph representation yields several benefits, e.g. efficient management and querying of networks, identification of subgraphs in networks, merging of SBGN diagrams/existing pathways into more extensive systems, or the comparison of different layers of granularity in SBGN languages.

Applications in biological and medicinal chemistry

The fields of Biology and Biochemistry have been a pioneer in the development of new data standards and knowledge representation paradigms, such as ontologies, to foster reuse, integration and translation of research data. These standards enable publicly available data resources such as UniProt (91), KEGG (92) and NCBI Taxonomy (93) to soft-link entities between each other, allowing the user to follow such links by manual browsing or through specialized workflows. The introduction of graph databases made it easier to integrate these resources explicitly. Built on Neo4j, Biochem4j (82) provides an integrated, queryable database that warehouses chemical, reaction, enzyme and taxonomic data from ChEBI (94), MNXref (95), Rhea

(96), KEGG, UniProt and the NCBI Taxonomy resources. Biochem4j translates ontology entities and raw biological data into an integrated graph representation, which, leveraged through Cypher query language, allows performing queries and detecting patterns across the whole range of available information.

Logically, graph representations apply to lower-level chemistry and related fields, like drug discovery research. One example is the fragment-based drug discovery (FBDD) (97), in which the validation stage of a project involves testing sensible close analogs of a fragment hit. This process needs adequate search tools to mine the many millions of similar compounds that are currently available in the fragment space from corporate collections or commercial suppliers. The Fragment Network (98) employs Neo4j to allow the user to search the chemical space around a compound of interest. The graph model treats each compound as a set of rings, linkers and substituents, with a resulting network containing a total of 23 million nodes and 107 million edges.

Applications in the omics domain

In the last five years, the usage of graph databases to support the integration of genomic, proteomic, metabolomic and phenotypic data has substantially increased. Most of the authors conclude that GDBMSs are valuable tools to deal with heterogeneity and lax structured data models because these provide a high degree of flexibility and lay the foundations for building integrated solutions.

Biological pathways

Repositories of metabolic maps, reconstructions, pathways and interactions provide fundamental tools for the biomedical investigation. Examples of these repositories are the Reactome Knowledgebase (99), Recon2 (100) and the latest development, Recon3D (101).

Reactome is a comprehensive repository of molecular reactions that include signal transduction, transport, DNA replication, protein synthesis and intermediary metabolism. Reactome contains a detailed representation of cellular processes, as an ordered network of molecular reactions, interconnecting terms to form a graph of biological knowledge. This structure serves both as an archive of biological processes and as a tool for discovering unexpected functional relationships in data. Reactome's data model initially follows a frame-based design stored in a relational MySQL database. Overcoming the relational model's intrinsic limitations requires an increased level of abstraction in its physical design to accommodate new concepts, ultimately affecting query complexity and execution time. As graph database systems have matured, the limitations of storing pathway data in relational databases have become more evident, motivating the project to develop tools to migrate the content into a Neo4j database (86, 87). The Reactome case is especially relevant because it exhibits a detailed description of the process to adopt a native graph database and how it improved the performance and capabilities of the whole system. On the one hand, the average query time dropped from 173.11 ms to 12.56 ms, a 93% reduction. On the other hand, the new graph model provides more straightforward ways to perform complex queries over metabolic pathways.

Recon2 is another large community-driven reconstruction of the human metabolic network, with thousands of reactions, unique metabolites and proteins, included in an SBML model. A model of this size and complexity comprises a challenge for advanced exploration involving associations between multiple concepts (e.g. network neighborhood of metabolites, shortest pathways between metabolites, proteins and complexes). Recon2Neo4j (76) is a Neo4j-based metabolic framework that models relevant concepts involved in the metabolic reactions as nodes in the graph database and the relationships among them as connecting edges, facilitating the exploration of comprehensive and highly connected human metabolic data and identification of metabolic subnetworks of interest.

HRGRN (80) is an integrative database for plant signal transduction, metabolism and gene regulation networks that is also backed by Neo4j. The solution, implemented as a web platform, provides the user with a graph-centered search interface to explore these biological systems, allowing to find potential paths or build either node-centralized or nodes-of-interest subnetworks. Regarding the data model, it followed an *ad hoc* approach, where biological entities (such as genes, proteins, small compounds and RNAs) are represented as nodes. For the relations between these entities, they defined eight types of edges that link the above nodes based on their biological functions. The Property Graph model is employed to attach a property indicating whether the relationship was validated or predicted.

BioGraphDB (102) is a bioinformatics database to combine different types of data from ten online public resources related to genes, microRNAs (miRNAs), proteins, pathways and diseases. To integrate these disparate resources, it builds on an Extract-Transform-Load (ETL) ecosystem capable of dealing with several formats (Tab delimited, XML, EBML and SQL) with a precise execution order to satisfy dependencies between the integrated resources. This process maps each biological entity and its properties into a vertex and its attributes, and relationships

between two biological entities into edges. In this case, the GDBMS of choice was OrientDB. When operating in graph mode, referenced relationships are like edges, accessible as first-class objects having start and end vertices and properties. This feature allows representing a relational model as a document-graph model, maintaining the relationships. With the end-user in mind, the Biograph web application (103) allows users to query, visualize and analyze biological data belonging to the sources available on BiographDB. However, the system is leaned toward a technical, graph knowledgeable audience, with explicit Gremlin query interfaces.

Similarly (104), Lysenko *et al.* illustrate how to build a graph structure to relate biomedical information at different levels and provide biological context to disease-related genes and proteins. It integrated genomic and proteomic data along with disease concepts to investigate possible relations between specific protein interactions, pathways, and typical phenotypes associated with asthma disease. In this case, the modeling strategy follows a protein-centric approach without a rigid schema or upper model (such as an ontology). This approach provides a higher degree of flexibility to integrate many semi-structured data sources and eases the development of *ad hoc* solutions, but at the expense of data standardization. The study provides a good insight into how graph databases can facilitate hypothesis generation. Another relevant contribution is to show how targeted Cypher queries exploit known structures, as well as graph algorithms like network neighborhood analysis, to provide biological context. An example of structural queries is obtaining proteins common to asthma and other related respiratory diseases, where protein nodes are connected to health conditions with a concrete 'associated' relation. They also demonstrate how simple graph traversal queries have the potential to assist in hypothesis generation by exploring relationships between concepts. For instance, to explore the relationship between asthma and alterations in circadian rhythm, they identify all shortest paths in the graph between asthma disease and a subset of protein-coding genes that generate and regulate circadian rhythms.

Epigenetics

Epigenetics is a growing area of research within the biomedical domain, and it is being used in many different contexts, such as the study of cancer. Existing relational databases that focus on various features of cancer pathways are restricted because the integration of multiple data types in relational databases is nontrivial, and the concept linking needed in the exploration of cancer-related hypotheses is limited. EpiGeNet (83) is a graph database that stores conditional relationships between molecular (genetic and epigenetic) events observed at different stages of colorectal oncogenesis. It integrates statistical data on molecular interdependencies recognized in colorectal cancer development, mined from StatEpigen (105) (a manually curated and annotated database) into a Neo4j instance. For the data model, 'MolecularEvent' nodes represent molecular events of conditional relationships, modeled as edges in the graph. The edge type is determined by phenotype information and the direction by the conditionality of the relationship. Attributes of 'MolecularEvent' are used to store event type and gene information, and the probability value is stored as a property of the edge. The resulting graph makes it possible to explore path connections associated with the highest 'incidence score' and employ Cypher queries in tasks like identifying genetic–epigenetic modifications, or molecular phenomena observed and reported in the specialized literature.

Transcriptomics

The transcriptome is the complete set of all RNA molecules in a cell, a population of cells, or in an organism (106). Transcriptomics studies generate large amounts of data, raw or processed, that may be deposited in public databases to make them available for a broader scientific community (107). These data can be expressed as gene expression and interaction networks, which may additionally be integrated with other biological datasets, such as protein–protein interactions (PPIs), transcription factors (TFs) and gene annotations. In this context and to evaluate the performance of Neo4j (46), Wiese *et al.* constructed Genome Regulatory Networks (GRNs) based on known enhancer–promoter interactions (EPIs) and their shared regulatory processes by focusing on cooperative TFs. Exploiting these data, we can find platforms like the non-coding RNA Human Interaction Database (ncRNA-DB), later evolved into Arena-Idb (79), miTALOS v2 (81), GeNNet (84), the Association Network Integration for Multiscale Analysis (ANIMA) (77) and the Gene Regulation Graph Database (GREG) (89). Except ncRNA-DB, all these platforms employ Neo4j as the GDBMS.

The ncRNA-DB is built on top of OrientDB, which translates class instances into nodes, permitting to follow an object-oriented design consisting of four main classes and its specializations: BioEntity, Alias, DataSource and Relation. The database imported and integrated associations among non-coding RNAs (miRNAs, circulating miRNAs, Long non-coding RNAs (lncRNAs) and other non-coding RNAs), genes, RNAs and associated diseases from 10 online databases. ncRNA-DB provides three alternative interfaces: a Cytoscape app named ncINetView, a web interface, and a command-line interface for raw resource

queries. Later, ncRNA-DB evolved into Arena-Idb, introducing several improvements like a mapping procedure for managing entities, an accurate integration process or reconstructed data storage. The updated dataset included seven new sources [such as Disease Ontology (108), lnc2cancer (109), lncACTdb (110), PSMIR (111), StarBase (112) or TarBase (113)]. Arena-Idb follows a hybrid RDBMS and GDBMS implementation by using MySQL to store names, annotations and sequences and Neo4j to handle the construction and visualization of the networks of thousands of biological entities.

To provide a tool to identify pathways regulated by miRNAs in a tissue-specific manner, miTALOS v2 employs Neo4j to integrate several heterogeneous data sources and directly model molecular entities and their interaction networks. This graph model represents miRNAs, genes, pathways and tissues as nodes. miRNAs are connected to genes with 'REGULATES' relationships, genes to tissues with 'EXPRESSED' and genes to pathways with 'MEMBER' relationships. The graph structure allows to, for instance, query the target genes of a miRNA expressed in a tissue or the pathways in which the target genes are involved. Furthermore, the schema-less approach enables the platform to keep updated and integrate new aspects like lncRNAs as regulators of gene expression or disease-specific expression profiles to extend tissue-specific gene expression.

GeNNet is an integrated transcriptome analysis platform that unifies scientific workflows with graph databases for selecting relevant genes according to the evaluated biological systems. The framework consists of three main components: the Scientific Workflow (GeNNet-Wf), the Graph database (GeNNet-DB) and the web interface (GeNNet-Web). GeNNet-DB uses an in-house data model to group nodes and edges into classes, according to the nature of the objects [e.g. GENE, BP (Biological Process), CLUSTER, EXPERIMENT and ORGANISM], and preloads a set of specified organisms to serve as the initial layout. Along with other associated elements, it includes genes annotated/described from ENTREZ (114) and their relationships integrated from STRING-DB (60), which contribute to posterior transcriptome analysis. The study provides analyses from the hepatocellular carcinoma (HCC) use case, demonstrating how concise graph operations through Cypher queries are capable of solving relatively complex topological questions, like finding the most connected genes that establish known connections to the PPI network. These genes act as hubs and may be associated with relevant pathways in the experimental context.

ANIMA allows the summarization and visualization of different views of the state of the immune system under different conditions and at multiple scales. The framework generates a multiscale association network from multiple data types by executing a comprehensive analytic workflow, enumerating bipartite graphs from the results and merging all graphs into a single network in Neo4j. ANIMA is architectural and conceptually similar to GeNNet, differing mainly in the detail of the implementation, the containerization approach, and the complexity of the model.

GREG is an integrative database that merges numerous source databases providing different scopes (e.g. DNA–DNA interaction, PPIs, bindings, DNA annotations or human cell data). It follows an in-house data model and takes advantage of the graph model to tackle challenges like integrating EPIs (with DNA binning strategy) or harmonizing data from chromatin interaction technologies with very different resolutions. When using small bins, its graph comprises more than 2 M nodes and more than 19 M edges, and the main limitation is that, due to include all non-coding regions, search time grows with the size of the genomic range. GREG provides both direct access to the Neo4j (via Cypher) and a friendly web platform. Through the web interface, the user can specify search parameters and access typical network analysis algorithms.

Biological knowledge graphs

While there exist multiple definitions of Knowledge Graphs (KGs) that depend on the application context (115), we can define them as large, heterogeneous knowledgebases modeled through graphs and ontologies, which derive new knowledge from existing datasets (116). KGs are undergoing a renewed interest not only in academia but in the industry as well (117). In addition to storing structured, contextual data, the principal reasons are the capability of obtaining new conclusions from existing data through reasoning (118), and the possibility to enrich machine-learning models by providing context and produce extra information through derived measures or embedding strategies (119–124). Lastly, advances in machine learning create new opportunities for automating the construction and exploitation of biological KGs (125). We summarize several platforms that, due to their broad integrative scopes, can be seen as Biological Knowledge Graphs.

The Monarch Initiative (85) is an ambitious endeavor that uses an ontology-based strategy to deeply integrate genotype–phenotype data from many species and sources, enabling computational interrogation of disease models and revealing complex genotype–phenotype relationships. Monarch employs RDF to ingest a variety of external data sources, modeling several complex data types and connecting entities from different databases. SciGraph (https://github.com/SciGraph/SciGraph) is its central database engine, which provides means to represent ontologies and

data described using ontologies as a Neo4j graph. The resulting combined corpus of graphs, from ontologies and ingested data, constitutes the Monarch Knowledge Graph. The platform provides several data access means for graph querying, application population and phenotype matching, as well as a web portal. The Monarch Web Portal (https://monarchinitiative.org/) exploits the graph to provide the users with several powerful features, in the likes of basic search, integrated information on entities of interest, search by phenotype profile, or text annotation.

Similarly, on a smaller scale, Pheno4J (75) provides a Java-based solution that loads annotated genetic variants and well-phenotyped patients into Neo4j. In order to build the database, Pheno4J requires user-generated files with the patient's genetic variant and phenotype relations on the one hand, and both the Human Phenotype Ontology (HPO) (126) and a gene-to-HPO file on the other.

Focused on the analysis and discovery of comorbid diseases in humans, GenCoNet (127) proposes a semi-automatic pipeline that provides the import, fusion and analysis of stable disease, gene, variant and drug data in a Neo4j database, resulting in a KG for network analysis of gene–disease associations. The workflow consists of four concrete steps. The first step determines comorbidities of high interest and obtains Disease Ontologies terms associated with genes. Secondly, the workflow obtains genes associated with disease variants from HPO, MalaCards (128), DisGeNet (129) and OMIM (https://omim.org). The third step determines the gene controlled by eQTL and associated with the disease. Lastly, it finds the drugs, extracted from DrugBank (130), which target genes and treats or contraindicate the disease. GenCoNet showcases the KG by employing network analysis to detect drug-induced diseases or contraindications of drugs.

We can also find hybrid approaches that utilize different database implementations to build the KG (131). Canevet *et al*. build on the Ondex software platform (132) and employ both triple stores and the Neo4j, which supports gene-evidence graph patterns by making the KGs accessible via Cypher. The data integration is harmonized through the Bio-Knowledge Network Ontology (BioKNO), a lightweight and general ontology. Likewise, focused on bacterial whole-genome sequencing (WGS), Spfy (88) employs ontologies and different database paradigms to integrate disparate data sources and formats. Spfy primarily uses Blazegraph (https://blazegraph.com/) for storage along with MongoDB (https://www.mongodb.com/) to cache a hash table for duplicate checking, arguing a more efficient approach than would be possible through a search of the graph structure. The graph allows retrospective comparisons across stored results as more genomes are sequenced or populations change.

As mentioned before, ontological and semantic approaches have proved its utility in knowledge-intensive domains like the biomedical domain. Exploiting semantic and logic descriptions is natural for graph databases and triple stores and can be of great importance in KG implementations. In contrast to the rest of similar efforts, BioGrakn (133) builds upon Grakn (https://grakn.ai/) to deliver a KG with deductive reasoning capabilities. It employs almost the same data sources as BioGraphDB, but its model is designed through an ontology implemented in Graql, the Grankn's declarative, knowledge-oriented graph query language. In the same vein as OWL and SWRL standards, Graql allows categorizing objects and relationships into distinct types, enabling inference and validation, used for searching genes linked to a particular Gene Ontology annotation, pathways linked to a particular gene, or finding all the upregulated differentially expressed (DE) miRNAs that also have validated mutations.

Discussion

The literature body shows several advantages when biomedical systems and applications employ a graph model in the storage layer. The graph model is especially useful for representing and accessing biological data because path-based queries are intuitive in biological networks, closer to real-world conceptualizations. RDF schema or OWL Bio-ontologies easily translate into a graph because they are already based on triples, which can be further expanded by identifying implied relations between classes through logical reasoning (134). Also, exploiting graph theory algorithms and subgraph matching queries enables the inspection and discovering patterns of interest within the graph structure. GDBMSs schema-less/schema-optional grants a high degree of flexibility in research settings, allowing applications to adapt and evolve quickly and introduce abstraction and specialization of entities and relations among them more easily. This adaptability eases data integration tasks, as we have seen in many of the integrative platforms.

Specialized, industry-ready GDBMSs are relatively new and well-established biological systems build upon conventional databases, typically RDBMSs. Relevant examples are the protein databases (135), which have to deal with millions of protein/complex interactions, as is PPI databases' case (136, 137). As described in the technical background, the underlying design of relational systems can lead to a trade-off between data integrity and performance.

BioGRID (138), for instance, approaches this problem by utilizing a suite of tables specifically engineered to optimize query time while maintaining a structured normalized form that does not compromise fundamental design principles. Other relevant databases like DIP (139), IntAct (140) and STRING (141) maintain their relational model to fulfill the storage needs without further considerations concerning performance.

As seen in section 2 (43), Have and Jensen employed STRING as the use case to evaluate GDBMSs in biomedical settings and confront against RDBMSs, generally finding better performance of the former in usual tasks in the context of PPI networks. In section 3, we also see that many applications integrate PPI databases by explicitly transforming protein entries as nodes and intermediate relationship tables directly as edges, reporting performance improvements with GDBMSs over RDBMSs in some of the reviewed works (45, 87). Still, it is important to remark that redesigning the data storage/access layer usually involves a notable development effort, which may discourage research teams (usually short in human and economic resources). Since most of the protein databases are freely available, it would be relevant to compare their current implementation and a GDBMS implementation through formal benchmarks in that specific scenario, justifying or not engaging such development.

There exist limitations and potential issues of which developers need to be aware. While ontologies avoid designing specific problem-oriented data models and minimize reliability issues, these may increment the model's complexity, jeopardizing the performance and integration time. If more relaxed schema approaches are adopted, the main trade-offs are deciding when certain data items become nodes or attributes and restraining both model complexity and integrity. Regarding performance, comparative benchmarks and more *ad hoc* studies are quite heterogeneous and show disparate findings in some cases, making it challenging to identify a performance baseline to favor a concrete technology. Those focused on specific problems, like biological questions, report better GDBMS performances and qualitative features for managing networks (42, 43, 45, 47, 52, 55, 56). More formal benchmarks (50) and (52) report superior RDBMS results in several categories, especially for grouping, sorting, aggregating and setting operations. However, in graph analytics workloads that mainly consist of multi-table joining, pattern matching or path identification, GDBMSs still perform better. The gap widens as the size of the dataset increases. Yet, some benchmarks report problems when the graph is large. In the case of Neo4j, the number of edges to evaluate and subgraph pattern matching size may be a performance pit. This situation requires GDBMSs to provide proper mechanisms, like node replication or partitioning, or forego features like schema-less as TigerGraph does. All in all, GDBMSs are not necessarily superior in all graph queries, and, like any development, the aims and operational context should dictate the technological choices.

From a development point of view, big projects naturally tend to adopt traditional relational databases because they require industry-level tools and libraries that ensure code quality and architectural features such as scalability, integration and standard design patterns. Both industry and communities back RDBMS implementations with reliable frameworks that ease its adoption with, for instance, database to object abstraction layers. However, at this point, many current GDBMS implementations also offer proper frameworks, programming interfaces and Object-Graph Mapping that fulfill such needs.

Another important consideration is the current lack of standardization of query languages and data access methods across GDBMS implementations at both syntactic and theoretical levels (142). Apache Tinkerpop (https://tinkerpop.apache.org/) provides a high-level framework and the functional graph traversal language Gremlin, but not all GDSMS integrate it and this approach implies more coupling with the application code. Neo4j's Cypher is a declarative language with similarities to common query languages and provides a clear graph path description syntax with full Create, Read, Update, Delete capabilities, making it one of the best solutions for graph querying. Cypher is the root of openCypher, a fully specified and open query language for property graph databases with >10 implementations across GDBMS solutions, even non-native ones like RedisGraph. TigerGraph follows a different approach with GSQL (https://docs.tigergraph.com/dev/gsql-ref), another powerful graph query language. It maintains backward compatibility with SQL, imposing a strict schema declaration in the query definition, and the queries behave as stored procedures, consisting of multiple SELECT clauses and imperative instructions such as branches and loops. This design targets enterprise applications, where the number and heterogeneity of external sources are not a concern, but instead, the size and performance, by optimizing storage format and query execution strategy, obtaining exciting results, as seen in Rusu and Huang (51). Fortunately, at the time of writing, the international committees that develop the SQL standard have voted to initiate *Graph Query Language* (GQL) (https://www.gqlstandards.org/) and intend to develop a declarative graph query language that builds on the foundations of SQL and integrates proven ideas from the existing openCypher, Oracle's PGQL, GSQL and G-CORE (143) languages, a move that ensures the future of GDBMSs.

We have seen different technologies come and go, and deciding a GDBMS that satisfy the necessities may also become a time-consuming task that can be seen as five steps or stages: problem analysis, requirements analysis, GDBMS analysis, benchmarking and GDBMS selection (144). Sites like https://db-engines.com provide useful ranks, comparative tables and insights that help in the selection process. From what we have seen in the literature, Neo4j outstands in its adoption, not only in the biomedical domain, mainly due to the powerful Cypher query language, decent performance and ease of implementation. We foresee that this situation will be less evident in the near future, given the number of competitive developments in the field.

Conclusion

In this work, we have followed the evolution and current landscape of GDBMSs, reviewed the bibliography looking for methods to evaluate their performance in different contexts and explored their applications in the biomedical domain. While RDBMSs and other NoSQL engines still provide better scalability options, more standardized query languages and more efficiency on typical data aggregation operations, most of the comparative analyses note that their performance suffers in densely connected datasets that imply a majority of many-to-many relations. Scenarios with a significant volume of complex relationships may benefit from GDBMSs for the following reasons: (i) graphs provide more natural modeling of many-to-many relationships; (ii) graph-oriented query languages provide more intuitive means for writing complex network traversal and graph algorithm queries than table-oriented ones like SQL, which require to join tables explicitly and reference columns; (iii) the schema-less/optional grants flexibility and (iv) in most situations, GDBMSs present higher performance for relationship-centric searches, like path traversals. These features yield several advantages for the biomedical domain, like easing the communication between domain experts, providing tools for discovering entities/clusters/patterns within the graph structure and facilitating data integration tasks, all of them very common when the investigation involves multiple sub-domains.

GDBMS technology is rapidly evolving to tackle scalability and similar operational weaknesses, offering a wide range of reliable choices to support the storage layer for either small prototypes or large, production-ready projects. The collection of described use cases and author experiences provides evidence that GDBMSs are very fit for biomedical data, as an individual storage system or as part of a hybrid, partitioned architecture. Moreover, by providing direct access to a graph model, late GDBMSs enable the use of graph algorithms and analytics in a very transparent way, improving hypothesis generation and testing.

Funding

This work was supported by the Norwegian Research Council (Dementia Disease Initiation) [grant number 217780]; Helse Sør-øst, NASATS (Dementia Disease Initiation) under [grant number 2013131]; Research project PID2019-110686RB-I00 of the State Research Program Oriented to the Challenges of Society.

Conflict of interest. None declared.

References

1. Bollobás,B. (1998) *Modern Graph Theory*. 1st edn. Vol. 184. Springer, New York, NY.
2. Harkins,S.S. and Reid,M.W.P. (2002). An introduction to relational database theory. In: *SQL: Access to SQL Server*. Apress, Berkeley, CA, pp. 35–68.
3. Codd,E.F. (1970) A relational model of data for large shared data banks. *Commun. ACM*, **13**, 377–387.
4. Hellerstein,J. and Stonebraker,M. (2005) *Readings in Database Systems*. 4th edn. MIT Press, Cambridge, MA.
5. Jamison,D.C. (2003) Structured Query Language (SQL) fundamentals. *Curr. Protoc. Bioinforma.*, **00**, 9.2.1–9.2.29.
6. Hsu,J.C., Hsu,C.H., Chen,S.C. *et al.* (2014) Correlation aware technique for SQL to NoSQL transformation. In: *2014 7th International Conference on Ubi-Media Computing and Workshops*, Institute of Electrical and Electronics Engineers Inc., Ulaanbaatar, Mongolia. pp. 43–46.
7. Singh,H. and Sharma,R. (2012) Role of adjacency matrix and adjacency list in graph theory. *Int. J. Comput. Technol.*, **3**, 179–183.
8. Liu,G., Chen,X., Wang,Z. *et al.* (2018) Evolving graph based power system EMS real time analysis framework. In: *IEEE International Symposium on Circuits and Systems*, May, IEEE, Florence, Italy. pp. 1–5.
9. Huang,H., Gao,Z., Dai,J. *et al.* (2020) Power grid modeling and topology analysis based on graph database conforming with CIM/E. In: Xue Y, Zheng Y, Rahman S (eds). *Lecture Notes in Electrical Engineering*. Vol. 585. Springer, Singapore, pp. 575–591.
10. Hu,K. and Zhu,J. (2018) A progressive web application on ancient roman empire coins and relevant historical figures with graph database. In: Ioannides M, Fink E, Brumana E, Patias P, Doulamis A, Martins J, Wallace M (eds). *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 11197 LNCS. Springer, Cham, pp. 235–241.
11. Kent,W. (1983) A simple guide to five normal forms in relational database theory. *Commun. ACM*, **26**, 120–125.
12. Chandrasekaran,B., Josephson,J.R. and Benjamins,V.R. (1999) What are ontologies, and why do we need them? *IEEE Intell. Syst. Their Appl.*, **14**, 20–26.
13. Konopka,B.M. (2015) Biomedical ontologies - a review. *Biocybern. Biomed. Eng.*, **35**, 75–86.

14. Smith,B., Ashburner,M., Rosse,C. *et al.* (2007) The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.*, **25**, 1251–1255.
15. Musen,M.A., Noy,N.F., Shah,N.H. *et al.* (2012) The National Center for Biomedical Ontology. *J. Am. Med. Informatics Assoc.*, **19**, 190–195.
16. Dovrolis,N., Stefanut,T., Dietze,S. *et al.* (2011) Semantic annotation and linking of medical educational resources. In: Jobbágy Á (ed.). *IFMBE Proceedings*. Vol. 37. Springer, Berlin Heidelberg, pp. 1400–1403.
17. Song,D., Chute,C.G. and Tao,C. (2012) Semantator: annotating clinical narratives with semantic web ontologies. *AMIA Jt. Summits Transl. Sci.*, **2012**, 20–209.
18. Shah,N. H., Bhatia,N., Jonquet,C., *et al.* (2009) Comparison of concept recognizers for building the open biomedical annotator. *BMC Bioinform.*, **10**.
19. El-Haj,M., Rutherford,N., Coole,M. *et al.* (2020) Infrastructure for semantic annotation in the genomics domain. In: *LREC*, Marseille.
20. Tan,H. and Lambrix,P. (2009) Selecting an ontology for biomedical text mining. In: *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, Association for Computational Linguistics, Boulder, Colorado, pp. 55–62.
21. Witte,R., Kappler,T. and Baker,C.J.O. (2007) Ontology design for biomedical text mining. In: Baker C. J. O, Cheung K-H (eds). *Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences*. Vol. 9780387484. Springer, Boston, MA, pp. 281–313.
22. Angles,R. and Gutierrez,C. (2008) Survey of graph database models. *ACM Comput. Surv.*, **40**, 1–39.
23. Angles,R. (2012) A comparison of current graph database models. In: *Proceedings - 2012 IEEE 28th International Conference on Data Engineering Workshops, ICDEW 2012*, IEEE, Arlington, VA, USA. pp. 171–177.
24. Buerli,M. and Obispo,C. (2012) The current state of graph databases. *Dep. Comput. Sci. Cal Poly San Luis Obispo, Calif.*, **32**, 1–7.
25. Miller,J.J. (2013) Graph database applications and concepts with Neo4j. In: *Proceedings of the Southern Association for Information Systems Conference*, Vol. 2324, AIS, Atlanta, GA, p. 36.
26. Kumar Kaliyar,R. (2015) Graph databases: a survey. In: *International Conference on Computing, Communication and Automation*, IEEE, Greater Noida, India. pp. 785–790.
27. Fernandes,D. and Bernardino,J. (2018) Graph databases comparison: allegrograph, arangoDB, infinitegraph, Neo4j, and orientDB. In: *DATA 2018 - Proceedings of the 7th International Conference on Data Science, Technology and Applications*, SciTePress, **Porto, Portugal**. pp. 373–380.
28. Roy-Hubara,N. and Sturm,A. (2020) Design methods for the new database era: a systematic literature review. *Softw. Syst. Model.*, **19**, 297–312.
29. Besta,M., Peter,E., Gerstenberger,R. *et al.* (2019) Demystifying graph databases: analysis and taxonomy of data organization, system designs, and graph queries.
30. Davoudian,A., Chen,L. and Liu,M. (2018) A survey on NoSQL stores. *ACM Comput. Surv.*, **51**.
31. Deutsch,A., Xu,Y., Wu,M. *et al.* (2019) TigerGraph: a native MPP graph database.
32. Cyganiak,R., Wood,D. and Lanthaler,M. (2014) *RDF*. RDF https://www.w3.org/TR/rdf11-concepts/.
33. Alaoui,K. (2019) A categorization of RDF triplestores. In: *Proceedings of the 4th International Conference on Smart City Applications - SCA'19*, ACM Press, New York, NY, USA. pp. 1–7.
34. Vilaça,R., Cruz,F. and Oliveira,R. (2010) On the expressiveness and trade-offs of large scale tuple stores. In: Meersman R, Dillon T, Herrero P (eds). *On the Move to Meaningful Internet Systems, OTM 2010*. Springer, Berlin Heidelberg, pp. 727–744.
35. Angles,R., Boncz,P., Larriba-Pey,J. *et al.* (2014) The linked data benchmark council: a graph and RDF industry benchmarking effort. *SIGMOD Rec.*, **43**, 27–31.
36. Iosup,A., Anderson,M., Tănase,I.G. *et al.* (2016) LDBC graphalyticsLDBC graphalytics: a benchmark for large-scale graph analysis on parallel and distributed platforms. In: *Proceedings of the VLDB Endowment*, Vol. 9, VLDB Endowment, pp. 1317–1328.
37. Cormen,T.H., Leiserson,C.E., Rivest,R.L. *et al.* (2009) *Introduction to Algorithms*. 3rd edn. MIT press, Cambridge, MA.
38. Page,L., Brin,S., Motwani,R. *et al.* (1999) *The PageRank Citation Ranking: Bringing Order to the Web*. Stanford InfoLab.
39. Gianinazzi,L., Kalvoda,P., De Palma,A. *et al.* (2018) Communication-avoiding parallel minimum cuts and connected components. In: *Proceedings of the 23rd ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming - PPoPP'18*, Vol. 53, ACM Press, New York, NY, pp. 219–232.
40. Boldi,P., Rosa,M., Santini,M. *et al.* (2011) Layered label propagation: a multiresolution coordinate-free ordering for compressing social networks. In: *Proceedings of the 20th International Conference on World Wide Web, WWW 2011*, Association for Computing Machinery, Hyderabad, India. pp. 587–596.
41. Schaeffer,S.E. (2007) Graph clustering. *Comput. Sci. Rev.*, **1**, 27–64.
42. Vicknair,C., Macias,M., Zhao,Z. *et al.* (2010) A comparison of a graph database and a relational database. In: *Proceedings of the 48th Annual Southeast Regional Conference on - ACM SE'10*, ACM Press, New York, NY, p. 1.
43. Have,C.T. and Jensen,L.J. (2013) Are graph databases ready for bioinformatics? *Bioinformatics*, **29**, 3107–3108.
44. Hoksza,D. and Jelinek,J. (2015) Using Neo4j for mining protein graphs: a case study. In: *2015 26th International Workshop on Database and Expert Systems Applications (DEXA)*, Vol. 2016-February, IEEE, Valencia, Spain. pp. 230–234.
45. Yoon,B.-H., Kim,S.-K. and Kim,S.-Y. (2017) Use of graph database for the integration of heterogeneous biological data. *Genomics Inform.*, **15**, 19.

46. Wiese,L., Wangmo,C., Steuernagel,L. *et al.* (2019) Construction and visualization of dynamic biological networks: benchmarking the Neo4j graph database. In: Auer S, Vidal M-E (eds). *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 11371 LNBI. Springer, Cham, pp. 33–43.
47. Batra,S. and Tyagi,C. (2012) Comparative analysis of relational and graph databases. *Int. J. Soft Comput. Eng.*, **2**, 509–512.
48. Angles,R., Prat-Pérez,A., Dominguez-Sal,D. *et al.* (2013) Benchmarking database systems for social network applications. In: *First International Workshop on Graph Data Management Experiences and Systems, GRADES 2013, co-located with SIGMOD/PODS 2013*. Association for Computing Machinery, New York, NY, USA. pp. 1–19.
49. Guia,J., Gonçalves Soares,V. and Bernardino,J. (2017) Graph databases: Neo4j analysis. In: *Proceedings of the 19th International Conference on Enterprise Information Systems*, Vol. 1, SCITEPRESS - Science and Technology Publications, Porto, Portugal. pp. 351–356.
50. Pacaci,A., Zhou,A., Lin,J. *et al.* (2017) Do we need specialized graph databases? In: *Proceedings of the Fifth International Workshop on Graph Data-management Experiences and Systems - GRADES'17*, ACM Press, New York, NY, pp. 1–7.
51. Rusu,F. and Huang,Z. (2019) In-depth benchmarking of graph database systems with the Linked Data Benchmark Council (LDBC) Social Network Benchmark (SNB).
52. Cheng,Y., Ding,P., Wang,T. *et al.* (2019) Which category is better: benchmarking relational and graph database management systems. *Data Sci. Eng.*, **4**, 309–322.
53. Erling,O., Averbuch,A., Larriba-Pey,J. *et al.* (2015) The LDBC social network benchmark: interactive workload. In: *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Vol. 2015-May, Association for Computing Machinery, Melbourne, VA, Australia. pp. 619–630.
54. Hurlburt,G.F., Thiruvathukal,G.K. and Lee,M.R. (2017) The graph database: jack of all trades or just not SQL? *IT Prof.*, **19**, 21–25.
55. Khan,W. and Shahzad,W. (2017) Predictive performance comparison analysis of relational and NoSQL graph databases. *Int. J. Adv. Comput. Sci. Appl.*, **8**, 523–530.
56. Khan,W., Ahmad,W., Luo,B. *et al.* (2019) SQL database with physical database tuning technique and NoSQL graph database comparisons. In: *Proceedings of 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference, ITNEC 2019*, Institute of Electrical and Electronics Engineers Inc, pp. 110–116.
57. Dominguez-Sal,D., Urbón-Bayes,P., Giménez-Vañó,A. *et al.* (2010) Survey of graph database performance on the HPC scalable graph analysis benchmark. *Lect. Notes Comput. Sci.*, **6185**, 37–48.
58. Chakrabarti,D., Zhan,Y. and Faloutsos,C. (2004) R-MAT: a recursive model for graph mining. In: *Proceedings of the 2004 SIAM International Conference on Data Mining*, Society for Industrial and Applied Mathematics, Philadelphia, PA, pp. 442–446.
59. Jouili,S. and Vansteenberghe,V. (2013) An empirical comparison of graph databases. In: *2013 International Conference on Social Computing*, IEEE, Alexandria, VA, USA. pp. 708–715.
60. Franceschini,A., Szklarczyk,D., Frankild,S. *et al.* (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.*, **41**, D808–D815.
61. Rodriguez,M.A. (2015) The Gremlin graph traversal machine and language (invited talk). In: *Proceedings of the 15th Symposium on Database Programming Languages - DBPL 2015*, Association for Computing Machinery, Pittsburgh, PA, pp. 1–10.
62. Cailliau,P., Davis,T., Gadepally,V. *et al.* (2019) RedisGraph GraphBLAS enabled graph database. In: *2019 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, IEEE, Rio de Janeiro, Brazil. pp. 285–286.
63. Welc,A., Raman,R., Wu,Z. *et al.* (2013) Graph analysis - do we have to reinvent the wheel? In: *1st International Workshop on Graph Data Management Experiences and Systems, GRADES 2013 - Co-located with SIGMOD/PODS 2013*, ACM Press, New York, NY, pp. 1–6.
64. Fan,J., Gerald,A., Raj,S. *et al.* (2015) The case against specialized graph analytics engines. In: *CIDR 2015-7th Biennial Conference on Innovative Data Systems Research* **CIDR**, **Asilomar**, **California**.
65. Zhao,K. and Yu,J.X. (2017) All-in-one: graph processing in RDBMSs revisited. In: *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Vol. Part F127746, Association for Computing Machinery, Chicago, Illinois, USA. pp. 1165–1180.
66. Xirogiannopoulos,K., Srinivas,V. and Deshpande,A. (2017) GraphGen: adaptive graph processing using relational databases. In: *5th International Workshop on Graph Data Management Experiences and Systems, GRADES 2017 - Co-located with SIGMOD/PODS 2017*, Vol. 7, Association for Computing Machinery, Inc, Chicago, IL, USA. p. 1–7.
67. O'Neil,P., Cheng,E., Gawlick,D. *et al.* (1996) The log-structured merge-tree (LSM-tree). *Acta Inform.*, **33**, 351–385.
68. Summer,G., Kelder,T., Ono,K. *et al.* (2015) cyNeo4j: connecting Neo4j and Cytoscape. *Bioinformatics*, **31**, 3868–3869.
69. Shannon,P., Markiel,A., Ozier,O. *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.
70. Saito,R., Smoot,M.E., Ono,K. *et al.* (2012) A travel guide to Cytoscape plugins. *Nat. Methods*, **9**, 1069–1076.
71. Hucka,M., Finney,A., Sauro,H.M. *et al.* (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, **19**, 524–531.

72. Lloyd,C.M., Halstead,M.D.B. and Nielsen,P.F. (2004) CellML: its future, present and past. *Prog. Biophys. Mol. Biol.*, **85**, 433–450.
73. Henkel,R., Wolkenhauer,O. and Waltemath,D. (2015) Combining computational models, semantic annotations and simulation experiments in a graph database. *Database*, **2015**.
74. Touré,V., Mazein,A., Waltemath,D. *et al.* (2016) STON: exploring biological pathways using the SBGN standard and graph databases. *BMC Bioinform.*, **17**, 494.
75. Mughal,S., Moghul,I., Yu,J. *et al.* (2017) Pheno4J: a gene to phenotype graph database. *Bioinformatics*, **33**, 3317–3319.
76. Balaur,I., Mazein,A., Saqi,M. *et al.* (2017) Recon2Neo4j: applying graph database technologies for managing comprehensive genome-scale networks. *Bioinformatics*, **33**, 1096–1098.
77. Deffur,A., Wilkinson,R.J., Mayosi,B.M. *et al.* (2018) ANIMA: association network integration for multiscale analysis. *Wellcome Open Res.*, **3**, 27.
78. Brandizi,M., Singh,A., Rawlings,C. *et al.* (2018) Getting the best of linked data and property graphs: Rdf2neo and the KnetMiner use case. In: *CEUR Workshop Proceedings*, Vol. 2275, CEUR-WS, Antwerp, Belgium.
79. Bonnici,V., Caro,G., De Constantino,G. *et al.* (2018) ArenaIdb: a platform to build human non-coding RNA interaction networks. *BMC Bioinform.*, **19**, 350.
80. Dai,X., Li,J., Liu,T. *et al.* (2016) HRGRN: a graph search-empowered integrative database of Arabidopsis signaling transduction, metabolism and gene regulation networks. *Plant Cell Physiol.*, **57**, e12.
81. Preusse,M., Theis,F.J. and Mueller,N.S. (2016) miTALOS v2: analyzing tissue specific microRNA function. *PLoS One*, **11**, e0151771.
82. Swainston,N., Batista-Navarro,R., Carbonell,P. *et al.* (2017) biochem4j: integrated and extensible biochemical knowledge through graph databases. *PLoS One*, **12**, e0179130.
83. Balaur,I., Saqi,M., Barat,A. *et al.* (2017) EpiGeNet: a graph database of interdependencies between genetic and epigenetic events in colorectal cancer. *J. Comput. Biol.*, **24**, 969–980.
84. Costa,R.L., Gadelha,L., Ribeiro-Alves,M. *et al.* (2017) GeNNet: an integrated platform for unifying scientific workflows and graph databases for transcriptome data analysis. *PeerJ*, **2017**.
85. Mungall,C.J., McMurry,J.A., Köhler,S. *et al.* (2017) The monarch initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res.*, **45**, D712–D722.
86. Fabregat,A., Jupe,S., Matthews,L. *et al.* (2018) The reactome pathway knowledgebase. *Nucleic Acids Res.*, **46**, D649–D655.
87. Fabregat,A., Korninger,F., Viteri,G. *et al.* (2018) Reactome graph database: efficient access to complex pathway data. *PLoS Comput. Biol.*, **2018**, 14.
88. Le,K.K., Whiteside,M.D., Hopkins,J.E. *et al.* (2018) Spfy: an integrated graph database for real-time prediction of bacterial phenotypes and downstream comparative analyses. *Database*, **2018**.
89. Mei,S., Huang,X., Xie,C. *et al.* (2020) GREG—studying transcriptional regulation using integrative graph databases. *Database*, **2020**, 1–8.
90. Le Novère,N., Hucka,M., Mi,H. *et al.* (2009) The systems biology graphical notation. *Nat. Biotechnol.*, **27**, 735–741.
91. Consortium,U. (2015) UniProt: a hub for protein information. *Nucleic Acids Res.*, **43**, D204–D212.
92. Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
93. Federhen,S. (2012) The NCBI taxonomy database. *Nucleic Acids Res.*, **40**, D136–D143.
94. Hastings,J., Owen,G., Dekker,A. *et al.* (2016) ChEBI in 2016: improved services and an expanding collection of metabolites. *Nucleic Acids Res.*, **44**, D1214–D1219.
95. Bernard,T., Bridge,A., Morgat,A. *et al.* (2014) Reconciliation of metabolites and biochemical reactions for metabolic networks. *Brief. Bioinform.*, **15**, 123–135.
96. Morgat,A., Lombardot,T., Axelsen,K.B. *et al.* (2016) Updates in rhea—an expert curated resource of biochemical reactions. *Nucleic Acids Res.*, **45**, D415–D418.
97. Erlanson,D.A., McDowell,R.S. and O'Brien,T. (2004) Fragment-based drug discovery. *J. Med. Chem.*, **47**, 3463–3482.
98. Hall,R.J., Murray,C.W. and Verdonk,M.L. (2017) The fragment network: a chemistry recommendation engine built using a graph database. *J. Med. Chem.*, **60**, 6440–6450.
99. Vastrik,I., D'Eustachio,P., Schmidt,E. *et al.* (2007) Reactome: a knowledge base of biologic pathways and processes. *Genome Biol.*, **8**, R39.
100. Thiele,I., Swainston,N., Fleming,R.M.T. *et al.* (2013) A community-driven global reconstruction of human metabolism. *Nat. Biotechnol.*, **31**, 419–425.
101. Brunk,E., Sahoo,S., Zielinski,D.C. *et al.* (2018) Recon3D enables a three-dimensional view of gene variation in human metabolism. *Nat. Biotechnol.*, **36**, 272–281.
102. Messina,A., Fiannaca,A., La Paglia,L. *et al.* (2018) BioGraph: a web application and a graph database for querying and analyzing bioinformatics resources. *BMC Syst. Biol.*, **12**, 98.
103. Messina,A., Fiannaca,A., la Paglia,L. *et al.* (2018) BioGraph: a web application and a graph database for querying and analyzing bioinformatics resources. *BMC Syst. Biol.*, **12**.
104. Lysenko,A., Roznovăţ,I.A., Saqi,M. *et al.* (2016) Representing and querying disease networks using graph databases. *BioData Min.*, **9**, 1–20.
105. Barat,A. and Ruskin,H.J. (2010) A manually curated novel knowledge management system for genetic and epigenetic molecular determinants of colon cancer. *Open Color. Cancer J.*, **3**, 36–46.

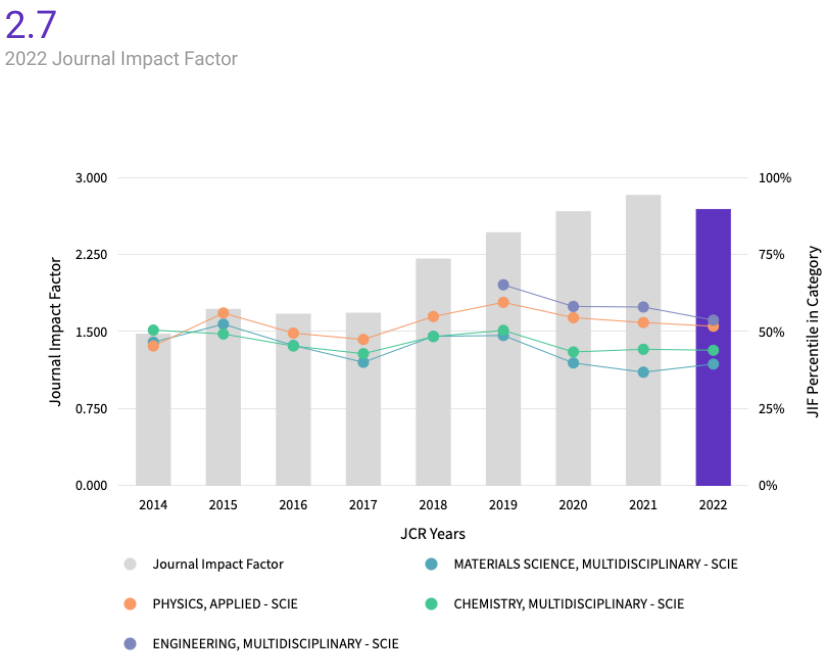
106. Blumenberg,M. (2019) Introductory chapter: transcriptome analysis. In: Miroslav B (ed). *Transcriptome Analysis*. IntechOpen, London, UK.
107. Lowe,R., Shirley,N., Bleackley,M. *et al.* (2017) Transcriptomics technologies. *PLoS Comput. Biol.*, **13**, e1005457.
108. Schriml,L.M., Arze,C., Nadendla,S. *et al.* (2012) Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Res.*, **40**, D940–D946.
109. Ning,S., Zhang,J., Wang,P. *et al.* (2016) Lnc2Cancer: a manually curated database of experimentally supported lncRNAs associated with various human cancers. *Nucleic Acids Res.*, **44**, D980–D985.
110. Wang,P., Ning,S., Zhang,Y. *et al.* (2015) Identification of lncRNA-associated competing triplets reveals global patterns and prognostic markers for cancer. *Nucleic Acids Res.*, **43**, 3478–3489.
111. Meng,F., Wang,J., Dai,E. *et al.* (2016) Psmir: a database of potential associations between small molecules and miRNAs. *Sci. Rep.*, **6**, 19264.
112. Yang,J.-H., Li,J.-H., Shao,P. *et al.* (2011) starBase: a database for exploring microRNA—mRNA interaction maps from Argonaute CLIP-Seq and Degradome-Seq data. *Nucleic Acids Res.*, **39**, D202–D209.
113. Sethupathy,P., Corda,B. and Hatzigeorgiou,A.G. (2006) TarBase: a comprehensive database of experimentally supported animal microRNA targets. *RNA*, **12**, 192–197.
114. Schuler,G.D., Epstein,J.A., Ohkawa,H. *et al.* (1996) Entrez: molecular biology database and retrieval system. *Meth. Enzymol.*, **266**, 141–162.
115. Sheth,A., Padhee,S. and Gyrard,A. (2019) Knowledge graphs and knowledge networks: the story in brief. *IEEE Internet Comput.*, **23**, 67–75.
116. Ehrlinger, L. and Wöß, W. (2016) Towards a definition of knowledge graphs. In: *CEUR Workshop Proceedings*, CEUR-WS, Vol. 1695.
117. Paulheim, H. (2017) Knowledge Graph Refinement: A Survey of Approaches and Evaluation Methods. *Semant. Web.*, **8**, 489–508
118. Chen,X., Jia,S. and Xiang,Y. (2020) A review: knowledge reasoning over knowledge graph. *Expert Syst. Appl.*, **141**, 112948.
119. Wang,Q., Mao,Z., Wang,B. *et al.* (2017) Knowledge graph embedding: a survey of approaches and applications. *IEEE Trans. Knowl. Data Eng.*, **29**, 2724–2743.
120. Grover,A. and Leskovec,J. (2016) Node2vec: scalable feature learning for networks. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Vol. 13, Association for Computing Machinery, San Francisco, CA, USA. 17 August 2016, pp. 855–864.
121. Xu,B., Liu,Y., Yu,S. *et al.* (2019) A network embedding model for pathogenic genes prediction by multi-path random walking on heterogeneous network. *BMC Med. Genomics*, **12**, 188.
122. Wang,X., Gong,Y., Yi,J. *et al.* (2019) Predicting gene-disease associations from the heterogeneous network using graph embedding. *Proceedings - 2019 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2019*, Institute of Electrical and Electronics Engineers Inc, San Diego, CA, USA. pp. 504–511.
123. Li,X., Chen,W., Chen,Y. *et al.* (2017) Network embedding-based representation learning for single cell RNA-seq data. *Nucleic Acids Res.*, **45**, e166.
124. Liu,X., Yang,Z., Sang,S. *et al.* (2019) Detection of protein complexes from multiple protein interaction networks using graph embedding. *Artif. Intell. Med.*, **96**, 107–115.
125. Nicholson,D.N. and Greene,C.S. (2020) Constructing knowledge graphs and their biomedical applications. *Comput. Struct. Biotechnol. J.*, **18**, 1414–1428.
126. Köhler,S., Doelken,S.C., Mungall,C.J. *et al.* (2014) The human phenotype ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res.*, **42**, D966–D974.
127. Shoshi,A., Hofestädt,R., Zolotareva,O. *et al.* (2018) GenCoNet – a graph database for the analysis of comorbidities by gene networks. *J. Integr. Bioinform.*, **15**, 1–9.
128. Rappaport,N., Twik,M., Plaschkes,I. *et al.* (2016) MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. *Nucleic Acids Res.*, **45**, D877–D887.
129. Piñero,J., Bravo,À., Queralt-Rosinach,N. *et al.* (2016) DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.*, **45**, D833–D839.
130. Wishart,D.S., Feunang,Y.D., Guo,A.C. *et al.* (2018) DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.*, **46**, D1074–D1082.
131. Brandizi,M., Singh,A., Rawlings,C. *et al.* (2018) Towards FAIRer biological knowledge networks using a hybrid linked data and graph database approach. *J. Integr. Bioinform.*, **15**, 3–4.
132. Canevet,C., Splendiani,A., Kuo,S.-C. *et al.* (2009) Ondex: data integration and visualisation for the semantic web. SWAT4LS, Vol. 559, RWTH Aachen University, Amsterdam.
133. Messina,A., Pribadi,H., Stichbury,J. *et al.* (2018) BioGrakn: a knowledge graph-based semantic database for biomedical sciences. In: Barolli L, Terzo O (eds). *Advances in Intelligent Systems and Computing*. Vol. 611. Springer, Cham, pp. 299–309.
134. Rodríguez-García,M.Á. and Hoehndorf,R. (2018) Inferring ontology graph structures using OWL reasoning. *BMC Bioinform.*, **19**, 7.
135. Chen,C., Huang,H. and Wu,C.H. (2017) Protein bioinformatics databases and resources. In: Wu CH, Arighi CN, Ross KE (eds). *Methods in Molecular Biology*. Vol. 1558. Humana Press Inc, New York, NY, pp. 3–39.
136. Ooi,H.S., Schneider,G., Chan,Y.L. *et al.* (2010) Databases of protein-protein interactions and complexes. *Methods Mol. Biol.*, **609**, 145–159.

137. Kangueane,P., Nilofer,C., Kangueane,P. *et al.* (2018) Databases for protein-protein interaction. In: *Protein-Protein and Domain-Domain Interactions*. Springer, Singapore, pp. 113–124.
138. Oughtred,R., Stark,C., Breitkreutz,B.J. *et al.* (2019) The BioGRID interaction database: 2019 update. *Nucleic Acids Res.*, **47**, D529–D541.
139. Salwinski,L., Miller,C.S., Smith,A.J. *et al.* (2004) The database of interacting proteins: 2004 update. *Nucleic Acids Res.*, **32**, D449–D451.
140. Aranda,B., Achuthan,P., Alam-Faruque,Y. *et al.* (2010) The IntAct molecular interaction database in 2010. *Nucleic Acids Res.*, **38**, D525–D531.
141. Szklarczyk,D., Morris,J.H., Cook,H. *et al.* (2017) The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.*, **45**, D362–D368.
142. Angles,R., Arenas,M., Barceló,P. *et al.* (2017) Foundations of modern query languages for graph databases. *ACM Comput. Surv.*, **50**, 1–40.
143. Angles,R., Arenas,M., Barceló,P. *et al.* (2018) G-CORE a core for future graph query languages. In: *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Association for Computing Machinery, Houston, TX. pp. 1421–1432.
144. Ortega,V., Ruiz,L., Gutierrez,L. *et al.* (2020) A selection process of graph databases based on business requirements. In: Mejia J, Muñoz M, Rocha A, Calvo-Manzano JA (eds). *Advances in Intelligent Systems and Computing*. Vol. 1071. Springer, Houston, TX, pp. 80–90.

Chapter 5

A Knowledge Graph Framework for Dementia Research Data

Title	*A Knowledge Graph Framework for Dementia Research Data*
Journal	Applied Sciences
Authors	Santiago Timón, Mariano Rincón, and Rafael Martínez Tomás, Bjørn-Eivind Kirsebom, and Tormod Fladby
Published	20 September 2023
Impact Factor	2.7
JCR Quartile	Q2
DOI	10.3390/app131810497



applied sciences

MDPI

Article

A Knowledge Graph Framework for Dementia Research Data

Santiago Timón-Reina [1,2,*], Mariano Rincón [3], Rafael Martínez-Tomás [3], Bjørn-Eivind Kirsebom [4,5] and Tormod Fladby [2,6]

[1] Escuela Internacional de Doctorado—Doctorado en Sistemas Inteligentes, Universidad Nacional de Educación a Distancia (UNED), 28015 Madrid, Spain
[2] Department of Neurology, Akershus University Hospital, 1478 Nordbyhagen, Norway
[3] Departamento de Inteligencia Artificial, Universidad Nacional de Educación a Distancia (UNED), 28015 Madrid, Spain
[4] Department of Neurology, University Hospital of North Norway, 9019 Tromsø, Norway
[5] Department of Psychology, Faculty of Health Sciences, UiT The Arctic University of Norway, 9019 Tromsø, Norway
[6] Institute of Clinical Medicine, Campus Ahus, University of Oslo, 0313 Oslo, Norway
* Correspondence: stimon3@alumno.uned.es

Featured Application: Applying knowledge graphs, graph analytics, and graph machine learning for integrating multi-modal dementia research data.

Abstract: Dementia disease research encompasses diverse data modalities, including advanced imaging, deep phenotyping, and multi-omics analysis. However, integrating these disparate data sources has historically posed a significant challenge, obstructing the unification and comprehensive analysis of collected information. In recent years, knowledge graphs have emerged as a powerful tool to address such integration issues by enabling the consolidation of heterogeneous data sources into a structured, interconnected network of knowledge. In this context, we introduce DemKG, an open-source framework designed to facilitate the construction of a knowledge graph integrating dementia research data, comprising three core components: a KG-builder that integrates diverse domain ontologies and data annotations, an extensions ontology providing necessary terms tailored for dementia research, and a versatile transformation module for incorporating study data. In contrast with other current solutions, our framework provides a stable foundation by leveraging established ontologies and community standards and simplifies study data integration while delivering solid ontology design patterns, broadening its usability. Furthermore, the modular approach of its components enhances flexibility and scalability. We showcase how DemKG might aid and improve multi-modal data investigations through a series of proof-of-concept scenarios focused on relevant Alzheimer's disease biomarkers.

Keywords: knowledge graphs; ontologies; graph databases; data modeling; dementia; omics

Citation: Timón-Reina, S.; Rincón, M.; Martínez-Tomás, R.; Kirsebom, B.-E.; Fladby, T. A Knowledge Graph Framework for Dementia Research Data. *Appl. Sci.* **2023**, *13*, 10497. https://doi.org/10.3390/app131810497

Academic Editors: José Ignacio Abreu Salas, Yoan Gutiérrez Vázquez and Ansel Yoan Rodríguez González

Received: 1 September 2023
Revised: 16 September 2023
Accepted: 18 September 2023
Published: 20 September 2023

1. Introduction

The dawn of "omics" technologies, accompanied by advancements in imaging, clinical data collection, laboratory testing, and phenotyping, has profoundly influenced biomedical research [1–7]. This multi-modal setting has provided an unprecedented, comprehensive view of complex biological systems, thereby inspiring a shift towards a more integrated understanding of diseases. However, the introduction of data from diverse modalities also presents unique challenges. Effectively integrating and interpreting the sheer volume, complexity, and diversity of data generated by these sources requires sophisticated computational tools. Moreover, the data, which are often distributed across various databases, publications, and repositories, pose considerable barriers to seamless data integration. Even more daunting is the task of transforming multi-modal data into clinically actionable insights, requiring the ability to connect data from molecular to clinical scales, a feat

complicated by the enormous diversity and complexity of individual diseases. These hurdles highlight the need for innovative strategies and tools to harness the potential of multi-modal data in propelling the field of precision medicine.

Since biological reality is often modeled as a network or graph [8,9], one technological approach that has gained significant traction is the use of knowledge graphs (KGs) [10], which allow for the integration and organization of diverse biomedical data types, facilitating their analysis and interpretation.

After Google introduced the knowledge graph in 2012, highlighting the advantages of the approach [11], KGs have become increasingly popular, finding adoption in industry with subsequent launches by companies such as Microsoft, Amazon, Airbnb, and Facebook [12], as well as in academia [13,14]. Nonetheless, the definition of KGs can vary based on the application context. In biomedicine, they can be characterized as data structures meant to gather and disseminate real-world knowledge, where nodes depict significant biomedical entities and the edges delineate diverse relationships that could exist between these entities [15]. KGs embody a methodological transition toward a more comprehensive representation of reality, facilitating the integration of heterogeneous data types and providing an intuitive, graph-based structure for representing intricate relationships between diverse biomedical entities.

Constructing a KG entails a series of methodological and technological decisions that profoundly impact the utility and effectiveness of the resulting product. A pivotal consideration in this process is the selection of a graph paradigm, which provides the theoretical and practical foundation for the structure and function of the KG. There are two primary approaches in this regard: Resource Description Framework (RDF) and Labeled Property Graphs [16–18]. Both of these approaches offer robust technological solutions, but each has its own strengths and weaknesses. While RDF offers standardization and robustness ideal for semantic applications, it may suffer from verbosity and computational inefficiency. Conversely, LPGs excel in their flexibility and intuitive structure, which allow for the straightforward representation of complex relationships and properties on both nodes and edges, but they may struggle in scenarios demanding high interoperability and standardization. Thus, the choice often hinges on the specific project requirements and constraints.

In addition to choosing a graph paradigm, selecting a data model or graph schema is another critical decision for building a KG. This model dictates how entities of interest and their relationships are represented within the KG. This aspect can be approached in two main ways: using an ad hoc data model tailored to the project's specific needs or adopting a standard model such as ontologies. In particular, biomedical ontologies have emerged as essential tools in standardizing terminology, modeling biological realities [19], supporting data annotation [20–23], and facilitating biomedical text mining [24,25]. With ongoing concerted efforts from the scientific community, these ontologies have evolved to incorporate fine-grained knowledge across various biomedical subdomains, as exemplified by initiatives such as the Open Biological and Biomedical Ontologies (OBO) Foundry [26] and the National Center for Biomedical Ontology (NCBO) [27] and its BioPortal [28]. Moreover, using logical modeling and annotation, biomedical ontologies make assertions that span and connect levels of biological organization, from the molecular level to phenotype and disease definitions. This ability to traverse and link multiple scales of biological information makes ontologies an invaluable resource for the construction of KGs for biomedical research.

The biomedical field is rich in open databases that offer scientific knowledge from various subdomains, including molecular biology (genomics, proteomics, and pathways), drugs, and disease characterization. These sources hold the potential for a more comprehensive understanding of biomedical phenomena; however, their value is often hindered by their dispersal across different platforms. KGs have emerged as instrumental tools for integrating and exploiting these disparate sources, fostering a multitude of projects that aim to unify the spread-out biomedical knowledge.

A prime example of such an initiative is the Monarch Initiative [29], which integrates genetic, phenotypic, and disease-related data to facilitate the identification of disease genes and variants. Similarly, the Clinical Knowledge Graph (CKG) [30] is an open-source platform that integrates proteomics, public databases, and literature. It effectively utilizes KGs to augment and enrich biomedical data, thereby facilitating informed clinical decision-making. Likewise, PrimeKG [31] is a multimodal KG that integrates a multitude of high-quality resources, representing various biological scales, i.e., from genotypes to clinical phenotypes. The scalable precision medicine open knowledge engine (SPOKE) [32] also integrates multiple biological data sources to provide structured knowledge ranging from low-level molecular biology to pharmacology and clinical practice. Furthermore, the KG-COVID-19 [33] project responded to the COVID-19 crisis by building a unified KG from disparate biomedical information about SARS-CoV-2, illustrating how KGs can effectively drive knowledge synthesis, particularly in emergent health situations.

As the number of available KGs increases, it has become evident that social and technical limitations exist, especially the need for standardization in entity naming and graph representation approaches [34,35]. Regarding modeling standardization, the Biolink Model [36] has emerged as a high-level data model that provides standard terms and relations for describing biological entities and their relationships for organizing data in biomedical KGs. Biolink serves both as a map for bringing together data from different sources under one unified model and as a bridge between ontological domains. As a similar initiative to OBO, centered around KGs, the KG-Hub project [37] provides a collection of tools and libraries for building interoperable KGs and a mechanism for sharing them to foster their reuse.

In addition to their ability to model and query data, graph analytics and graph machine learning techniques have made notable advancements [38,39], supported by open-source libraries such as GRAPE [40] and KGTK [41]. One technique particularly relevant in the biomedical domain is graph embedding [42–47], which allows us to capture complex graph structures into lower-dimension vectors. Exploiting these features to integrate specific patient data with large biomedical KGs has already shown promising results in deriving actionable clinical outputs, as evidenced by advancements in understanding diseases such as multiple sclerosis [42] and Alzheimer's disease [48]. Recent dementia research uses multi-modal data to understand the condition from various aspects, including genomics, transcriptomics, metabolites, imaging, and clinical features. Having a framework that enables the systematic construction and instantiation of research and clinical data in a standardized manner offers significant benefits.

This paper introduces DemKG, a KG framework designed specifically for dementia research needs. The framework leverages reference ontologies from OBO, standard KG technologies from KG-Hub, and an instantiation tooling to transform source data into the KG following sound design patterns within the ontological model. DemKG reuses most of its knowledge sources, provides specific terminological extensions to cover gaps identified in the scope of dementia, and ingests biological databases of interest, resulting in an integrative KG that covers the multiple data modalities involved in the research, including genomics, proteomics, imaging, fine-grained phenotyping, and clinical tests. Thanks to its design, DemKG is easily extensible, delivering means to customize and deploy in modern graph databases for enhanced data querying and retrieval. The expressive knowledge model supports advanced analytics through graph and network algorithms, which play an active role in the progression of research and better patient care through the implementation of precision medicine.

2. Related Work

Advancements in storage and graph technologies, coupled with the increasing availability of open scientific data, have led to the emergence of multiple biological KGs [49]. Projects such as the Monarch Integrated Knowledge Graph, the Clinical Knowledge Graph

(CKG), PrimeKG, and the scalable precision medicine open knowledge engine (SPOKE), previously introduced in the introduction, bear similarities to our initiative.

The Monarch Integrated Knowledge Graph [29] is a notable example of biological KGs, which assimilates various data types (including genotype, phenotype, and disease) from multiple sources into a unified semantic graph model. The Monarch KG has been instrumental in our project, DemKG, as it not only serves as a primary data source but also offers an array of tools we utilize. Our philosophy aligns closely with that of the Monarch KG, emphasizing a robust semantic foundation while integrating data from a variety of external sources, including other ontologies and extensions. We build upon this work to extend it with dementia-related knowledge and provide means for integrating study data.

CKG [30] is an open-source platform designed to harmonize a wide range of "omics" data types into a coherent structure, including genomics, transcriptomics, proteomics, and metabolomics. CKG favors a custom data model formed from a selected set of concepts and relationships from specific ontologies. On top of the KG, CKG integrates statistical and machine learning algorithms to streamline the analysis and interpretation of typical proteomics workflows. DemKG resonates with CKG's mission to improve the modeling and integration of omics data. However, it deviates fundamentally from its approach to data modeling, wherein CKG employs a more circumscribed model.

PrimeKG [31] is a multimodal KG for precision medicine analyses. Like its counterparts, it integrates a plethora of resources to describe a broad spectrum of diseases with relationships across major biological scales. One of them is combining the entire range of approved drugs with their therapeutic action, distinguishing it from other systems. Moreover, unlike DemKG, PrimeKG employs a custom approach to its data model, incorporating ten types of nodes and thirty types of undirected edges extracted from reference ontologies. Furthermore, it lacks a systematic schema to integrate experimental and study data.

SPOKE [32] is a KG that connects information from 41 biological data sources, structured as 21 different node types and 55 edge types, ranging from low-level molecular biology to pharmacology and clinical practice. It uses 11 different ontologies to organize the data semantically meaningfully and, in its last iteration, also integrates the Biolink model whenever it is found to be practical. SPOKE is implemented as a Neo4j database built from a collection of Python scripts and provides a graphical user interface and a REST API for end-user access. Our method stands distinct from SPOKE in several crucial aspects. Primarily, it offers an open toolkit for KG construction and personalization, ensuring both platform and representational paradigm autonomy. Moreover, despite utilizing a comparable modeling approach, DemKG fosters a closer connection with a vast array of domain ontologies by preserving links to explicitly defined terms and relationships. Finally, our framework provides a flexible and robust module for research data integration.

In summary, our work distinguishes itself from similar efforts through a comprehensive approach that integrates a well-established terminological foundation and community standards, follows design patterns conducive to data integration, and defines terminological extensions specific to the dementia domain, facilitated through a dedicated low-code solution for seamless study data integration.

3. Materials and Methods

3.1. Terminological Foundation

In the construction of the knowledge graph, the initial and pivotal decision revolves around selecting an appropriate graph schema to provide a solid conceptual base that effectively captures data entities drawn from the array of biological subdomains pertinent to dementia research. This choice presents a dichotomy: one option involves creating a flexible, ad hoc schema tailored to the identified needs, while the alternative entails adopting a more structured strategy that employs standard terminologies and ontologies. Our methodology aligns with the latter approach, and a fundamental design principle in the construction of our KG is the utilization of domain reference ontologies to ensure the following:

1. The concept definitions are concise, accurate, and relevant;
2. There exists an active community keeping the ontology updated;
3. They are widely recognized, cross-referenced, and follow consistent design patterns.

The criteria set forth are congruent with the guiding principles of the OBO foundry. OBO endorses an extensive range of domain-specific ontologies that are distinguished by well-demarcated scopes, the reutilization of concepts across ontologies, and alignment with a unified upper-level model, specifically the Basic Formal Ontology (BFO) [50], and relations are defined in the Relations Ontology (RO). Given these attributes, we gave preferential consideration to OBO ontologies during our selection process.

As the KG must cater to a variety of domains, adopting this approach enables us to concentrate mainly on integration and only define new terms when detecting a gap. Some notable examples of the employed OBO ontologies include the Gene Ontology [51,52], Chemical Entities of Biological Interest (CHEBI) [53], and Protein Ontology (PR) [54] for the genetic and molecular domain. For the phenotype and disease domain, we utilize the Monarch Disease Ontology (MONDO) [55], Human Phenotype Ontology (HP) [56,57], and Phenotype And Trait Ontology (PATO) [58]. In the area of anatomy, we incorporate the Uber-Anatomy Ontology (UBERON) [59,60] and the Foundational Model of Anatomy (FMA) [61]. For neuropsychological tests and their relations, we include the Neuropsychological Testing Ontology (NPT) [62] and the Neurocognitive Integrated Ontology (NIO) [63]. For modeling experimental settings, the Ontology for Biomedical Investigations (OBI) [64,65] plays a central role.

These ontologies provide a significant level of detail, and reusing or referencing concepts between them expands the knowledge network, facilitating the exploitation of multi-domain and multi-level relations. For example, this interconnectedness simplifies navigation from HP phenotypes referenced in a disease definition in MONDO to specific genes in GO, proteins in PR, and molecular entities in CHEBI. Furthermore, we also include relevant Monarch data and annotation ingestions; specifically, gene and gene-phenotype annotations, filtered protein–protein interactions from the STRING database [66], and pathway knowledge from the Reactome pathway knowledgebase [67]. The complete list of knowledge sources and annotations is listed in Table 1.

Table 1. List of DemKG knowledge sources.

Source	Source Identifier	Reference
Basic Formal Ontology	BFO	[50]
Biolink model	biolink	[36]
Chemical Entities of Biological Interest	CHEBI	[53]
Cell Ontology	CL	[68]
Evidence and Conclusion Ontology	ECO	[69]
Environmental Factor Ontology	EFO	[70]
Gene Ontology	GO	[52]
Gene Ontology Annotations	GOA	-
Human Phenotype Ontology	HP	[57]
Human Phenotype Ontology Annotations	HPOA	-
Information Artifact Ontology	IAO	-

Table 1. *Cont.*

Source	Source Identifier	Reference
Mass Spectrometry Ontology	MS	[71]
Mondo Disease Ontology	MONDO	[55]
Monarch KG	Monarch	[29]
Neurocognitive Integrated Ontology	NIO	[63]
Neuropsychological Testing Ontology	NPT	[62]
Ontology of Biological Attributes	OBA	[72]
Ontology for Biomedical Investigations	OBI	[65]
Ontology for General Medical Science	OGMS	[73]
Ontology of Medically Related Social Entities	OMRSE	[74]
Phenotype And Trait Ontology	PATO	[58]
Phenomics Integrated Ontology	PHENIO	-
Protein Ontology	PR	[54]
Relations Ontology	RO	-
Reactome	Reactome	[67]
Scientific Evidence and Provenance Information Ontology	SEPIO	-
STRING database ingestion	STRING	[66]
Uber Anatomy Ontology	UBERON	[59]

While the standardization offered by domain ontologies is undoubtedly a strength, it can also impose limitations due to the inherent trade-off with flexibility. This high level of detail can complicate the integration of non-OBO ontologies and external datasets. Additionally, querying the graph requires a comprehensive understanding of the underlying model. We employ the Biolink model as our high-level data model to mitigate these issues. Biolink offers a means to utilize higher-level concepts from its "category" hierarchy while still allowing references to more specific ontology terms. The same versatility is available for relationships through the use of the "related_to" hierarchy, thus providing a balance between standardization and flexibility in our knowledge graph.

3.2. Terminological Extensions

OBO covers most of the conceptualization needs, but gaps remain relevant to the implementation. To overcome this issue, we implement an application ontology that is also one of the inputs of the merging process. The primary interventions relate to phenotypic normality, as well as to the necessary assay and platform definitions missing from OBI.

HP and MONDO thoroughly model disease states, conditions, and abnormal phenotypes, leaving out any reference to normal counterparts. To allow the categorization of instances of normal/healthy cases, we introduced a "Phenotypic normality" hierarchy. This new hierarchy is modeled as a sibling branch of the HP "Phenotypic abnormality", mirroring its hierarchy to allocate the "normality" concepts of interests.

In dementia research, the utilization of neuropsychological assessments such as the Mini-Mental State Examination (MMSE) [75], the Consortium to Establish a Registry for Alzheimer's Disease (CERAD) wordlist memory test (WLT) [76], Visual Object and Space Perception (VOSP) battery [77], Trail Making Test (TMT) [78], Clock Drawing Test [79], and Controlled Oral Word Association Test (COWAT-FAS) [80] is instrumental in quantifying cognitive function domains and tracking disease progression. We have implemented the necessary concepts to cover CERAD, VOSP, and COWAT-FAS tests, with the primary classes

allocated under the "cognitive function assay" branch of NPT, while also relating to the mental and cognitive functions they assess.

The AT(N) classification system [81] is another tool of great importance for assessing the subject's biological state and understanding the intricate relationships between key biomarkers and their impact on disease evolution. AT(N) categorizes biomarkers according to their role in the disease progression, namely, Beta-amyloid deposition (A), pathologic tau (T), and neurodegeneration (N). Within each biomarker category, values can be positive or negative (+/−), derived from defined normal or abnormal cut points, resulting in the creation of eight distinct AT(N) "biomarker profiles" (Table 2). To provide proper terminological coverage, we have defined new classes for each biomarker profile and phenotype terms related to abnormal CSF protein concentration phenotypes related to phosphorylated tau (P-tau) and total tau (T-tau) missing from HP. Each biomarker profile is defined under the "value specification" class from OBI, with asserted logical axioms to associate them with the specific phenotype.

Table 2. AT(N) biomarker profiles and categories as defined by the NIA-AA Research Framework. Each biomarker profile is modeled as a descendant of the "value specification" class defined in OBI.

AT(N) Profiles	Biomarker Category	
A-T-(N)-	Normal AD biomarkers	Alzheimer's continuum
A-+T-(N)-	Alzheimer's pathologic change	
A+T+(N)-	Alzheimer's disease	
A+T+(N)+	Alzheimer's disease	
A+T-(N)+	Alzheimer's and concomitant suspected non-Alzheimer's pathologic change	
A-T+(N)-	Non-AD pathologic change	
A-T-(N)+	Non-AD pathologic change	
A-T+(N)+	Non-AD pathologic change	

3.3. Technical Implementation

The implementation consists of three main software pieces covering different parts of the KG generation, integrated into a building pipeline: the extensions ontology builder, the KG-builder, and the data transformer module. To maximize effectiveness and reproducibility, in all three sub-projects, we employ state-of-the-art ontology and graph tooling maintained by the community and relevant projects such as Monarch and the "universal biomedical data translator" from the National Center for Advancing Translational Sciences (NCATS) [82].

The extensions ontology builder produces an OWL ontology using the Ontology Development Kit (ODK) v1.4.1 [83] as the building framework. The ODK provides a pre-configured, standardized environment with a set of tools that support all stages of the ontology lifecycle (creation and editing, building, and testing, and releasing with version control) and ensures a systematic approach to ontology maintenance. When possible, we define new classes that follow a pattern using the Dead Simple OWL Design Patterns (DOS-DP) v0.1.10 [84], reducing manual editing and consequently reducing errors and improving reproducibility. All the axioms are kept under OWL2 [85] DL profile.

The KG-builder is responsible for obtaining the different sources of knowledge and merging them into the terminological KG. Built upon the KG-Hub tooling ecosystem, the main configuration inputs are the merge and download YAML descriptor files, guiding the download and merge steps. When available, the ontologies are downloaded from the KG-Hub repository [86]. OBO ontologies are already maintained as Biolink-compliant graphs in the Knowledge eXchange Format (KGX) [87] in the KG-OBO project [88] and are directly merged from each specific release artifact. The merging step includes all downloaded sources and the extensions ontology to obtain a final KGX graph.

One challenge when converting OWL ontologies into a graph structure lies in the difficulty of accessing class relationships established through subclass and class equivalence axioms. These assertions hold significant value in capturing the biomedical knowledge outlined in the comprehensive OBO ontologies. To address this situation, both the ontology and builder modules materialize class equivalence axioms. In the context of the extensions ontology, we utilize the relation-graph [89] library during the later stages of the construction process. In the case of OBO ontologies, the KG-builder retrieves a subset of links from the materialization output within Ubergraph [90], which also employs relation-graph.

The transformer module is a Python solution that provides an accessible approach to generating graph data in KGX format from tabular source input. This module adopts a YAML-based transform definition schema, mirroring the approach of other tools in the pipeline. This schema adheres to a standardized structure wherein users can define mappings from columns to specific classes paired with various instantiation design patterns. The schema effectively models common research entities, including medical history, physical examination, and measurement assays, all aligned with dedicated instantiation patterns that are further elaborated upon in the subsequent subsection.

The builder pipeline integrates all steps and can be configured to generate two artifacts: solely the terminological graph or the terminological graph with data instantiation.

3.4. Data Transformation Design Patterns

One of the aims of the KG is to integrate raw research data to enable explicit connections with knowledge concepts. We propose a set of design patterns to support the data instantiation of patient/subject study visits, phenotype observations arising from these visits, measurements/analyses derived from samples collected from different specimens, and neuropsychological test results. In all these patterns, OBI is the central ontology employed to enable the relating of clinical and research concepts with specific entities of the biomedical domain. Figures 1–3 illustrate the main patterns through simplified concept map figures, depicting the main ontology classes and properties involved, identified with a pseudo-CURIE of the format PREFIX, namely, "class label", where prefix is the OBO ontology prefix.

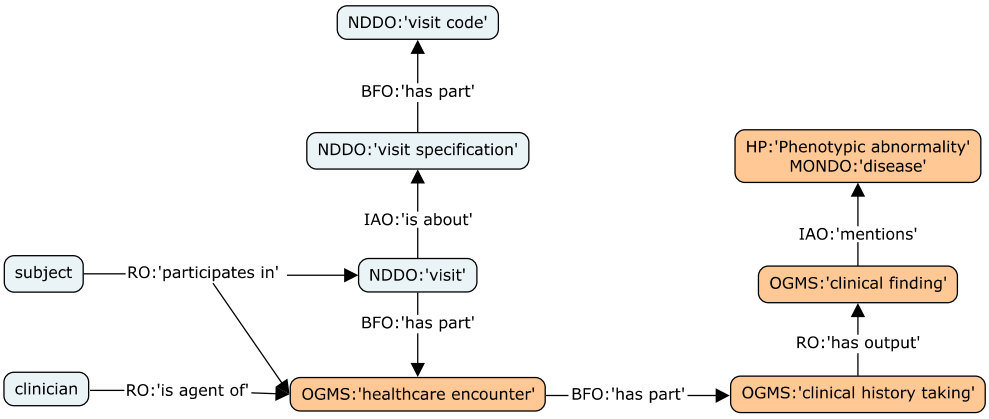


Figure 1. Concept map of the visits (light blue) and clinical (orange) design patterns, depicting the main ontology classes employed to model data entities.

The first pattern models the relations between study protocol/visit encounters, the agents involved, and the resulting outputs. The pattern mainly utilizes concepts defined in the Neurodegenerative Disease Data Ontology (NDDO) [91] (integrated in NIO) and the Ontology for General Medical Science (OGMS). The pattern supports a proper logical definition of longitudinal protocols, common in dementia research studies.

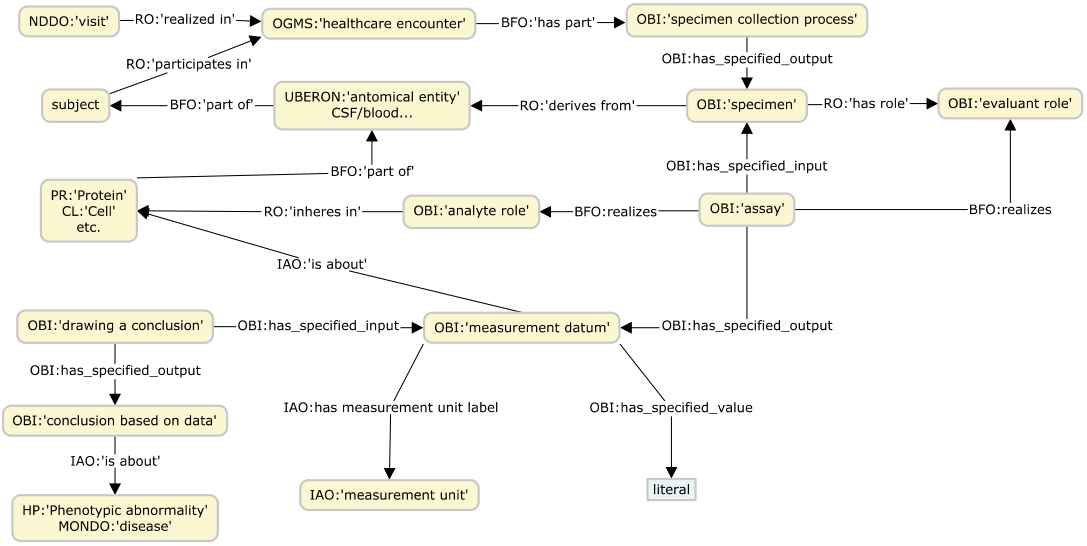


Figure 2. Concept map of the experimental measurements design pattern.

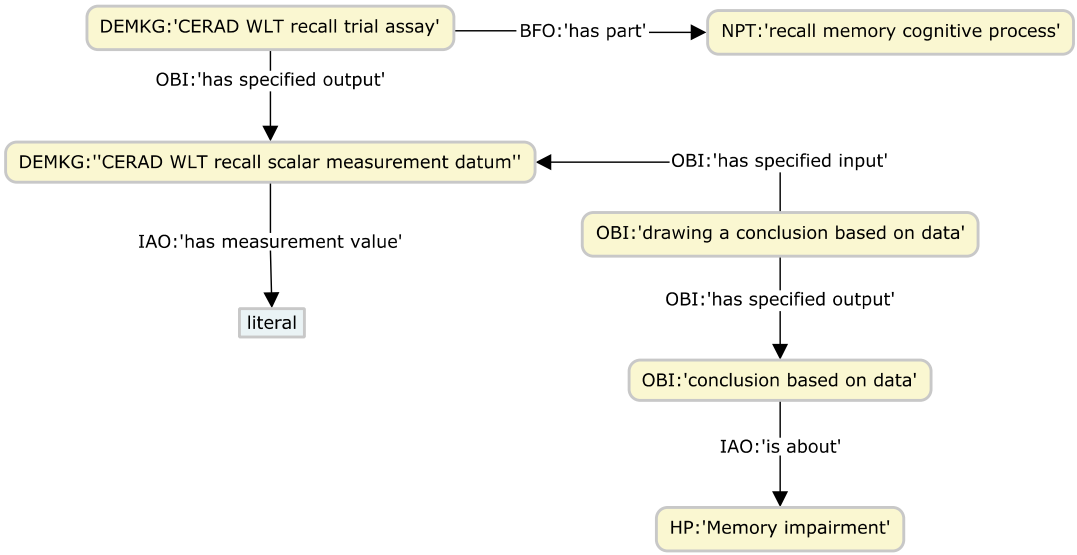


Figure 3. Exemplification of the neuropsychological test design pattern, through a CERAD recall test.

Clinical history phenotypes are characterized through observations at a study visit or from existing records. The framework leverages a pattern that relates visits with specific clinical administration, the finding, and the observed phenotype, usually a phenotype or disease concept from MONDO or HP. Relevant metadata can also be linked to the OGMS clinical entities, such as dates, agents involved, and locations. This pattern is shared across medical history, physical examination, and diagnosis processes. Figure 1 illustrates both the visits and clinical patterns.

A critical component of research data encompasses various assay measurements and proteomic datasets. We employ OBI's assay design patterns [92] to capture the multiple aspects involved in this process. These patterns enable the comprehensive integration of data pertaining to the assay, the specimen, and the molecule or material under examination, such as a protein or leukocyte count. Several relevant ontologies, including GO, PR, and Cell Ontology (CL), supply the necessary terminologies. We leverage entities from UBERON to denote the anatomical origin of the sample. This pattern facilitates the preservation of crucial metadata about processes, encompassing information about the type of assay, the specimen or sample employed, experimental conditions such as freeze–thaw cycles, and the date and time of collection. Such metadata is of considerable value for resource

management and can significantly aid research analyses. For instance, the type of tube in which a sample was collected could influence assay results and should be accounted for in linear models. Overall, it provides a more comprehensive context of the conditions under which experiments are conducted, enhancing the reproducibility and reliability of experimental outcomes.

Analyses derived from neuroimaging techniques, including segmentation measurements from tools such as Freesurfer [93] and Automatic Sub Hippocampal Segmentations (ASHS) [94], along with white matter evaluations from Diffusion Tensor Imaging (DTI) [95] and peak width of skeletonized mean diffusivity (PSMD) [96], play an indispensable role in dementia research. The pattern supporting this data modality follows a similar approach to the previous one, illustrated in Figure 2. To associate the measured anatomical entities, we utilize the FMA, which offers precise terms to align with the parcellation regions delineated by the widely used brain atlases in segmentation software, particularly for hemisphere-specific terms. More general terms from UBERON can be obtained using the "xref" property, employed for mapping concepts between different ontologies.

The last design pattern focuses on effectively relating the information content of a given test with the cognitive domain, providing means by which to stratify subjects via cognitive staging and the specific domain or phenotypic abnormality from HP at query time. This pattern exploits the axioms that connect cognitive tests with the evaluated domains.

4. Results

We have developed a KG framework that harmonizes biomedical knowledge and evidence from various sources, coupled with a transformation module designed to streamline the integration of multi-modal and omics data in dementia research. The core components of the framework encompass the extensions ontology builder, which provide ontological definitions to fill identified gaps from the domain ontologies; the KG-builder, in charge of obtaining, merging, and producing the KG; and the data transformer module, a low-code interface to transform source study data. All components are publicly accessible on GitHub (https://github.com/demkg-framework/, accessed on 30 August 2023). This trio of tools forms an intuitive building pipeline and also offers flexibility for customization, enabling users to construct the graph from scratch, adapt it to specific requirements, and deploy it on their preferred platform and graph database.

The backbone of our implementation is rooted in established community standards, technologies, and methodologies. The initial step involved the selection of a comprehensive array of domain reference biomedical ontologies, primarily from OBO, to form an expressive knowledge model for our primary KG. These ontologies offer a variety of well-defined concepts across varying levels of granularity, encapsulating intricate details of biological reality in the form of hierarchical relationships and concept networks.

To facilitate a consistent term mapping across various ontologies and mitigate computational demands, we utilized pre-built KGs from the KG-Hub initiative and the KG-OBO subset as our foundation, employing the KGX tool for the merging phase of the KG-builder pipeline. The KG-Hub initiative utilizes the Biolink model as its high-level data model, which we adopted to introduce greater flexibility and provide a comprehensive yet adaptable terminology overlay on the ontological model. The Biolink model facilitated the creation of both relaxed and detailed modeling and query capabilities, thereby enhancing the standardization and flexibility of our model. The default KG consists of 1.5 M nodes and 11.5 M edges.

To fill the identified gaps in the foundational model, we developed specific terminological extensions through the extensions ontology. We employed ODK to systematically introduce new terms, leveraging the OBO ecosystem to import and extend relevant external terms using DOS-DP whenever feasible.

Finally, the transformation module provides a low-code solution to transform tabular source data and generate necessary instance nodes and edges by following specific

design patterns that effectively depict study visits, phenotype observations, measurements/analyses derived from samples, and neuropsychological test results. These design patterns promote efficient data instantiation under the ontological model of the source research data, interconnecting various aspects of the study design outputs and providing a robust platform for data querying and network-oriented analyses. Figure 4 shows an overview of the framework components.

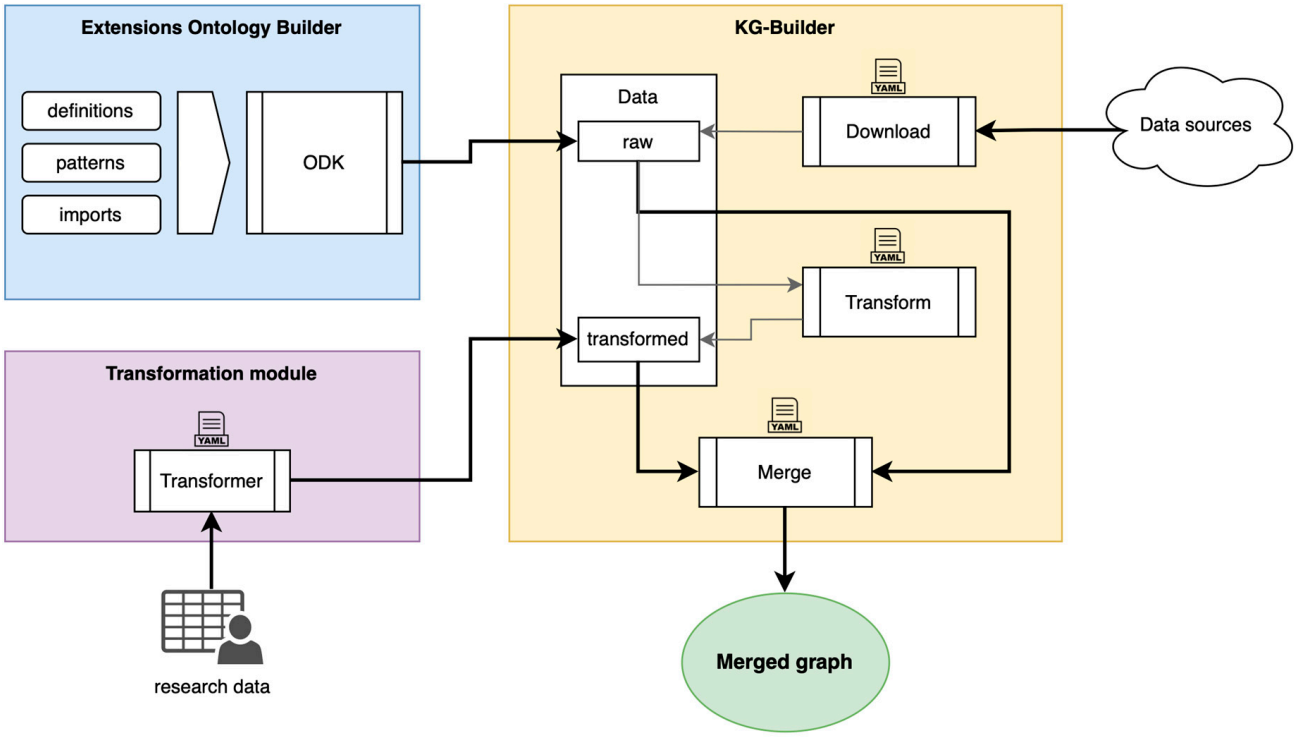


Figure 4. Overview of the DemKG framework components.

4.1. Use Case: Graph-Enabled Phenotype, Flow, and Protein Exploration from AT(N) Biomarker Profiles

To validate the DemKG framework, we applied it to the Dementia Disease Initiation (DDI) study data, a multi-site longitudinal observational study aimed at identifying early biomarkers for patients at risk of developing dementia [97]. The DDI dataset encompasses a range of clinical items, including medical history, standardized physical, neurological, and cognitive examinations, as well as laboratory and proteomic assays derived from blood and cerebrospinal fluid (CSF) samples, MRI, FDG-PET, and amyloid PET imaging, along with genomic analyses. We integrated these diverse data modalities and explored various aspects of the key biomarkers of the AD continuum, as categorized by the AT(N) classification.

4.1.1. Experimental Setup

The central DDI data platform is the XNAT archiving system [98], which is complemented by tailored customizations and data export functionalities, including automatic biomarker-based AT(N) classification, and population-adjusted norming for pertinent screening tests such as CERAD [99,100], VOSP [77], and TMT [78,101]. We implemented the transformation descriptor for the DDI data, involving direct mappings from clinical codes and rules to translate assay and experiment results into specific phenotype and disease entities. We then fed the descriptor along with the aggregated Comma-separated values (CSV) dump from XNAT to the transformation module to obtain the graph representation.

The DDI cohort graph comprises 96,939 nodes and 362,824 edges, whereas an average subject subgraph with four visits has 3469 nodes and 8284 edges. This transformed graph was merged into the final DDI-KG, which we ingested using the KGX module into a Neo4j Community instance deployed in a Podman container configured with eight cores and 16 GB of RAM, running on the secured servers of the TSD (Tjeneste for Sensitive Data) facilities managed by the University of Oslo. We opted for Neo4j due to its widespread adoption, the capabilities of its Cypher query language, and its reliable performance. Furthermore, KGX automatically creates node indices and constraints to improve loading and query performance for this platform.

Taking advantage of these features, the setup proves efficient with the resultant graph model, particularly for queries with clearly defined traversals and designated node labels. Figure 5 offers a preliminary analysis for estimating query performance, tracing the time consumed in navigating paths that extend from one to ten hops from subject nodes to various relevant node types in the graph. As anticipated, the number of target nodes considerably affects query performance, primarily driven by the increased number of edges to evaluate and traverse, coupled with the augmented data volume to handle. This scenario is especially pronounced in the most populated and interconnected node types, namely, proteins, genes, and diseases. Therefore, queries involving numerous or unrestricted quantities of such nodes require thoughtful design.

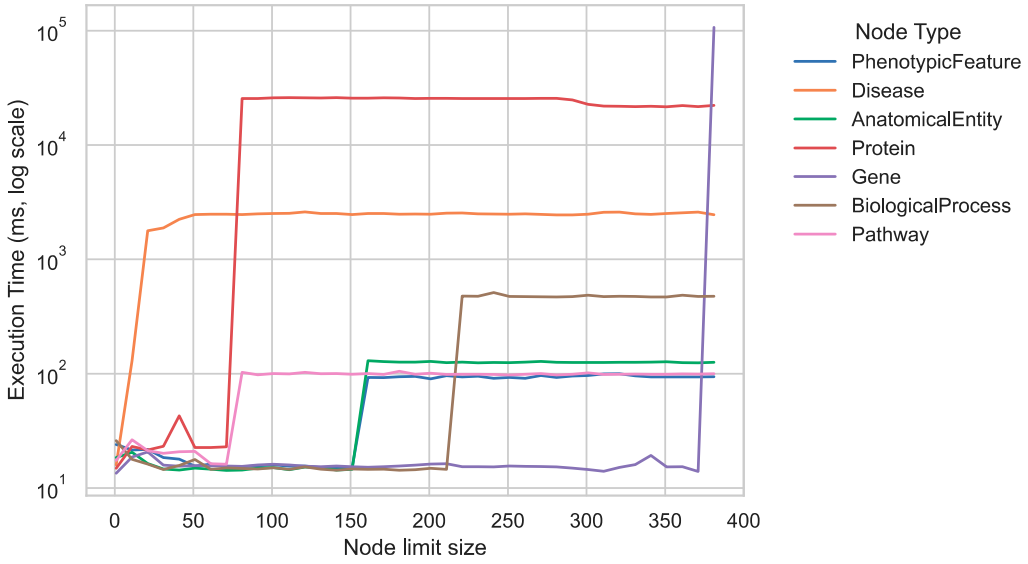


Figure 5. Mean execution times over ten runs for variable-length traversal queries between 1 and 10 connections, navigating from subject nodes to key Biolink categories.

4.1.2. Experimental Results

A key objective of the DDI study is to comprehend the evolution of subjects across different disease states within the biological reality, and the AT(N) classification system is a pivotal reference point. The developed design patterns facilitate connections at various levels, enabling the exploration of individual and group trajectories across visits and expediting the retrieval of relevant phenotypes using graph queries (Figure 6).

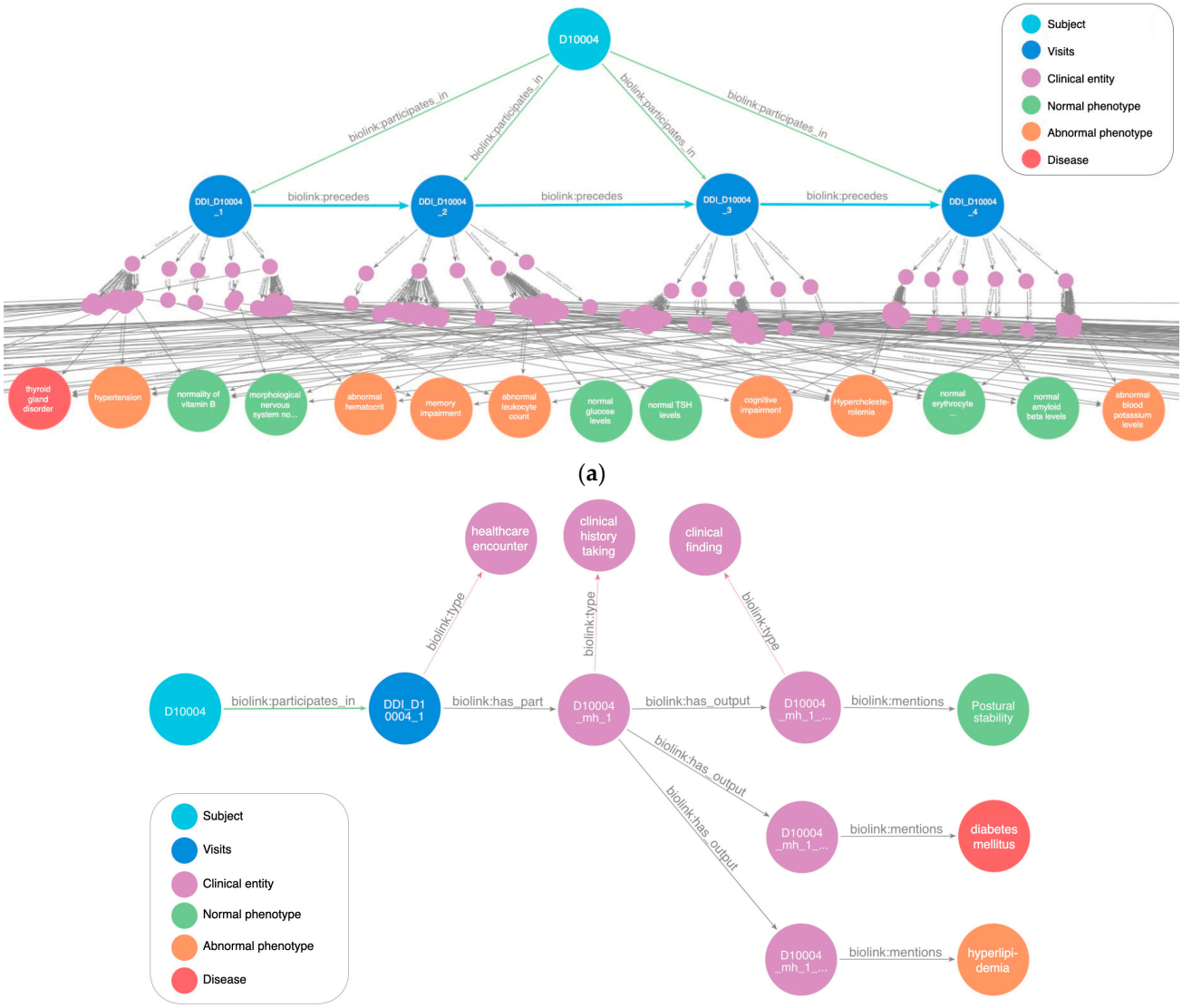
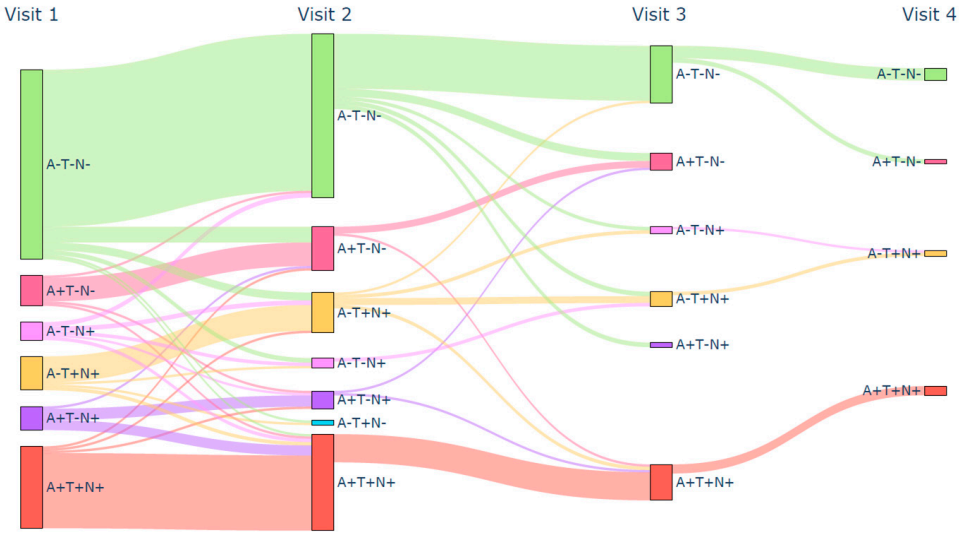


Figure 6. A DDI subject subgraph that illustrates study visits and associated phenotypes, visualized with Neo4j Bloom and further edited for readability. (**a**) An overview of longitudinal visits. Subjects are connected to each visit via the "biolink:participates_in" predicate. The logical sequencing of visits is established through the "biolink:precedes" predicate, facilitating query traversal. Clinical entity nodes represent associated medical processes (medical history, cognitive screenings, lab assays, and more), serving as the source of observations and conclusions while also supplying context and metadata for encounters and experimental setups. These nodes link to phenotype and disease entities to depict the outcomes of the clinical/research processes. (**b**) A specific visit branch tracing the path from the individual subject to the evaluated phenotypes and diseases noted during a medical history recording. Additional data from clinical entities are omitted to maintain clarity and uphold subject privacy.

Using the AT(N) entities defined in the extensions ontology, we queried the graph database to investigate the flow between the different biomarker profiles. This exploration helped unravel the transitions between them at the cohort level, aiding in data filtering for parallel research endeavors. Moreover, presented visually (Figure 7), the outcomes of these queries proved instrumental in quality control efforts by highlighting unlikely transitions from pathological to normal states. Such interventions are vital since AT(N) profiles derive

from biomarker measurements, where unexpected transitions may result from issues or errors in the respective assays.



(**a**)

```
// AT(N) biomarker profiles are subclass of AT(N) value specifications in the extension ontology
MATCH
    (atn_val)-[:`biolink:subclass_of`]->(:`biolink:NamedThing` {id:'DEMKG:1900000'})
// source AT(N) values
MATCH
    (subject:`biolink:IndividualOrganism`)-[:`biolink:participates_in`]->(v1:`biolink:ClinicalEntity`)
    -[:`biolink:has_part`]->()-[:`biolink:has_output`]->()-[:`biolink:has_part`]->(atn_val)
// target ATNs values are collected from follow-up visits, using 'precedes' predicate
MATCH
    (v1)-[:`biolink:precedes`]->(vtarget:`biolink:ClinicalEntity`)-[:`biolink:has_part`]
    ->()-[:`biolink:has_output`]->()-[:`biolink:has_part`]->(target_atn_val)
// Prepare source AT(N), target AT(N), and count
WITH
    atn_val.name + '_' + v1.visit AS source,
    target_atn_val.name + '_' + vtarget.visit AS target,
    count(DISTINCT subject.name) AS count
WHERE count > 1
RETURN  source, target, count
ORDER BY source, target
```

(**b**)

Figure 7. Graph-based analysis illustrating the transitional flow among AT(N) biomarker profiles within the DDI cohort over successive protocol visits. (**a**) A Sankey diagram depicting the transitions in biomarker profiles. (**b**) The Cypher query utilized to calculate transition counts based on the predefined AT(N) biomarker profiles in the ontology.

As shown in Figure 7b, one of the valuable attributes of KGs that incorporate domain ontologies is richer semantic querying. Leveraging the hierarchical structure within phenotype and disease ontologies, we exploited semantic querying to gather phenotypes spanning different domains and visualized their prevalence across the AT(N) profiles. As depicted in Figure 8, we focused on phenotypes extracted from the "Abnormality of higher mental function" class within the HP ontology. Phenotypes related to memory, language, and executive function were referenced based on the rules established for the norming items in the cognitive screening section of the dataset descriptor.

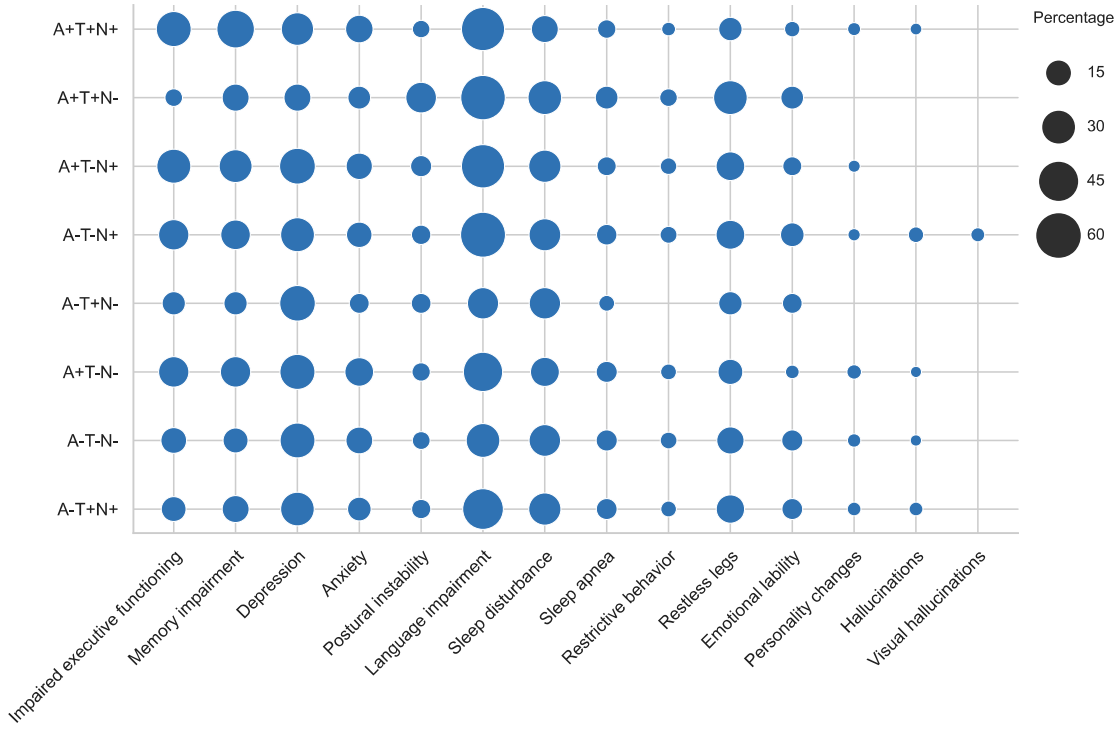


Figure 8. Dot plot from the collected phenotypes from subjects and their prevalence among the different AT(N) biomarker profiles.

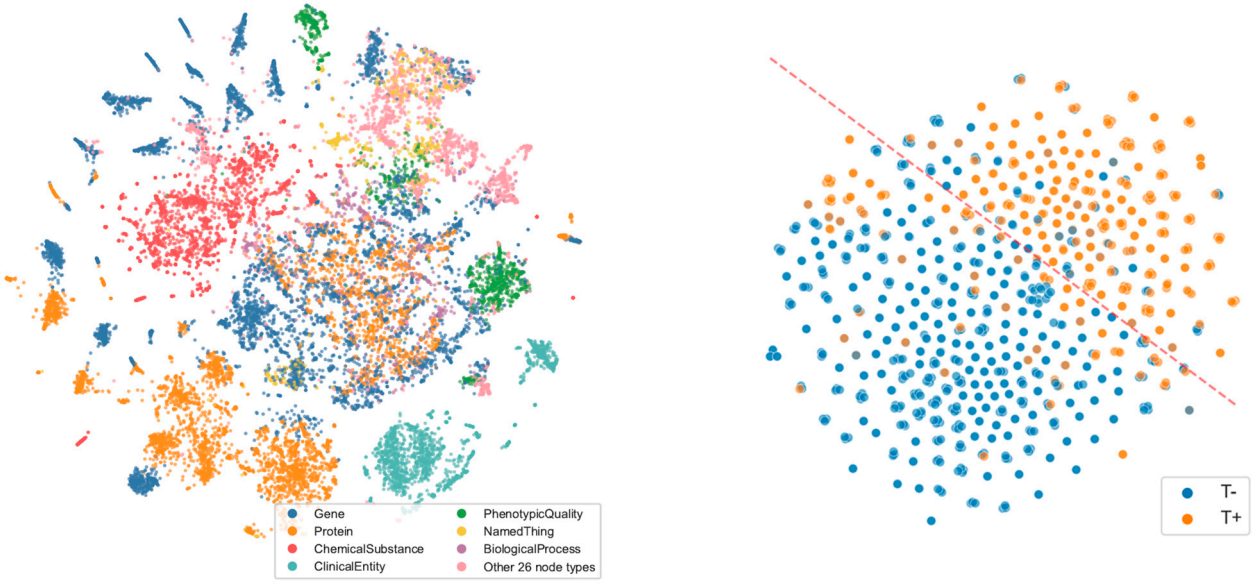
To capture complex graph structures into low-dimensional vector space, we utilized the GRAPE library to create node embeddings using the node2Vec algorithm [102] with Skip Gram [103] and applied them to evaluate various aspects of the AT(N) biomarkers.

We conducted an interesting experiment to investigate if the embeddings of subject visits showed any patterns in the low-dimensional space or were influenced by specific AT(N) profiles. Using t-SNE [104] to reduce the embeddings to two dimensions, we observed a clear tendency for Tau pathology to group together in the embedding space, suggesting shared characteristics among the phenotypes assessed in those visits. The visit node embeddings are visualized in Figure 9, accompanied by a decision boundary computed through a logistic regression model.

Lastly, we combined the graph query capabilities, node embeddings, and topological metrics to obtain a broader overview of the relationships between assay proteins and the AT(N) protein biomarkers to assist in decision-making processes that could steer future analyses. Since the graph provides explicit links between available assays and the analytes being evaluated, we gathered CSF-derived ELISA and proteomics target proteins for comparison, focusing on the shared network encompassing GO biological processes (BPs).

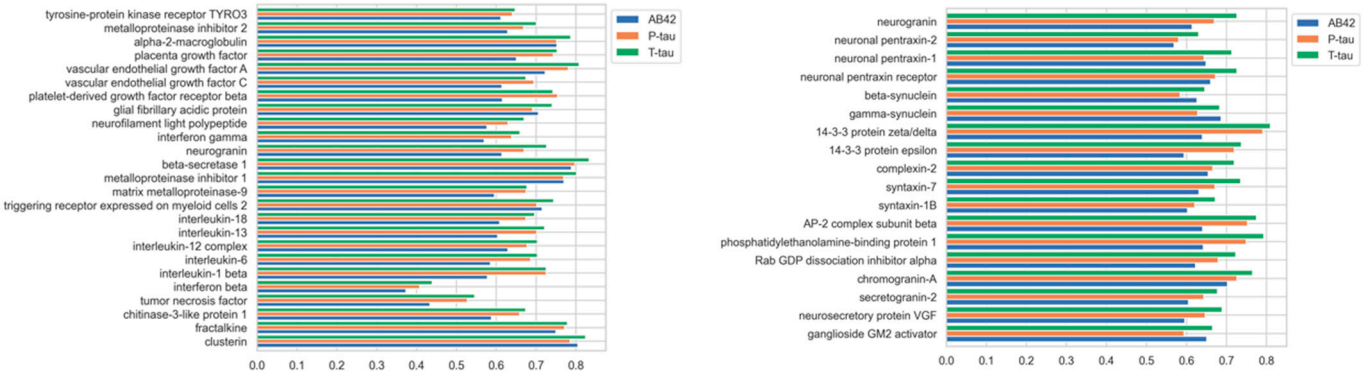
For assessing protein relationships, we employed a simple pair-wise cosine similarity measure. This allowed us to quickly gauge how closely protein nodes were related and then rank the proteins that were most closely associated with the AT(N) panel (Figure 10).

To examine shared BPs between AT(N) and the assessed proteins, we employed a graph query to obtain the extensive network of protein activities. Given that proteins participate in thousands of such processes, to enhance navigability, we used GRAPE to calculate node betweenness and closeness centrality metrics, utilizing them as indicators of node relevance for prioritizing and narrowing down the pool of BPs to be investigated. A snapshot of this process is depicted in Figure 11.



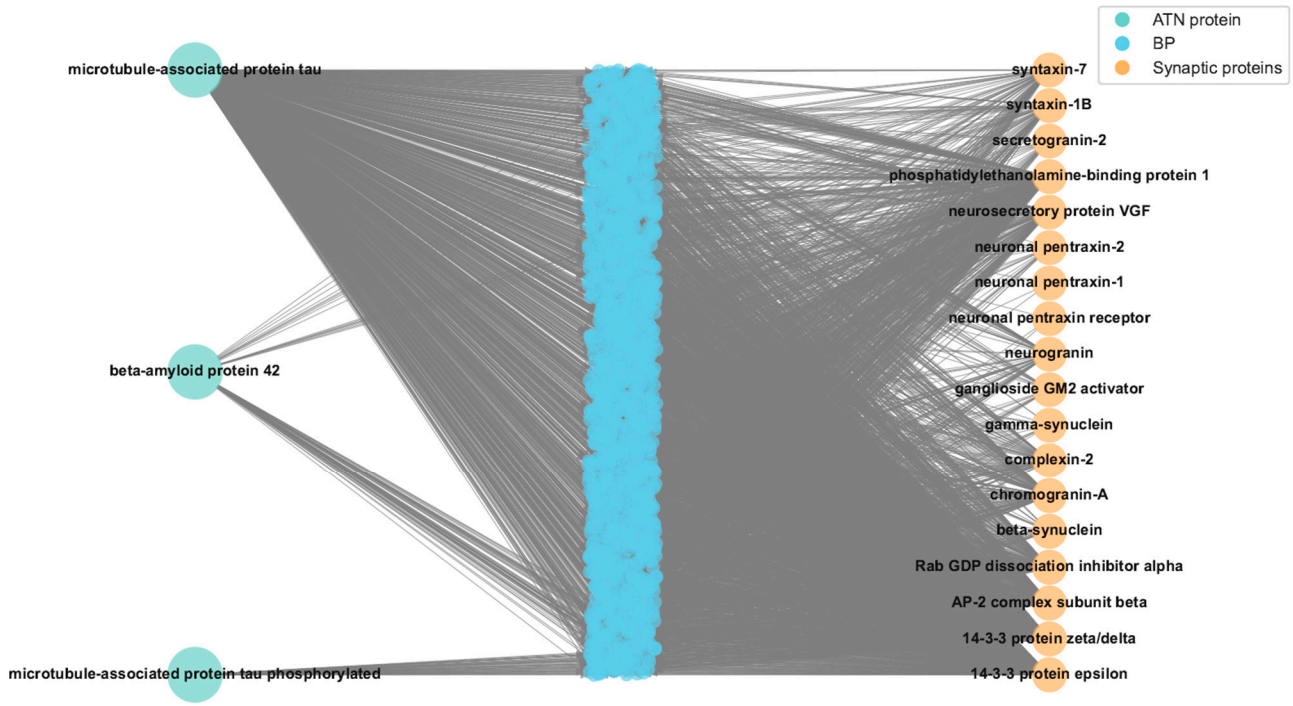
(**a**) (**b**)

Figure 9. t-SNE visualizations of node embeddings. (**a**) Scatter plot output from GRAPE for all node embeddings from the KG representing the topological connectivity, colored by node type. It displays similarity and some possible clusters (Balanced accuracy: 60.32% ± 1.25%); separability consideration derives from evaluating a Decision Tree trained on five Monte Carlo holdouts, with a 70/30 split between training and test sets. (**b**) Visit node embeddings with nodes labeled by their associated T biomarker from AT(N) (pathologic tau). The dashed line marks the decision boundary between node types computed from a logistic regression model, with an accuracy of 0.831.

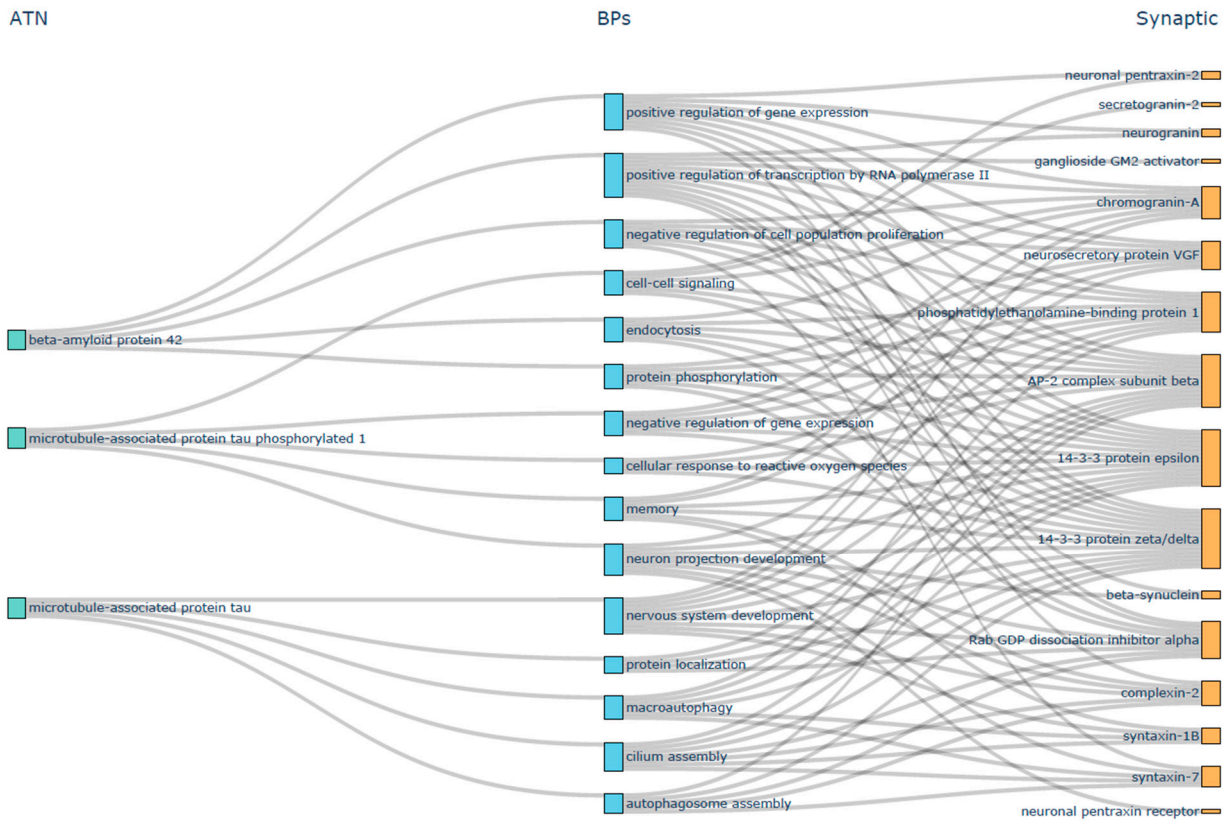


(**a**) (**b**)

Figure 10. Cosine similarity of target proteins to AT(N) proteins. (**a**) CSF ELISA protein panel. (**b**) Synaptic protein panel from proteomics assays.



(**a**)



(**b**)

Figure 11. A snapshot of BP prioritization from node centrality. (**a**) Full subnetwork of shared BPs between AT(N) and synaptic panel proteins. (**b**) Sankey diagram with the top 10 BPs obtained from closeness centrality.

5. Discussion

In our work, we introduce DemKG, a KG framework designed to integrate various ontologies and knowledge sources to focus on dementia research data. This framework aims to cover terminological and design needs for multi-modal and omics data, with additional terminological extensions developed when necessary. We also followed specific patterns to cater to typical dementia research data outputs.

A key advantage of DemKG is its flexibility and ease of extension or customization to adapt to particular needs, made possible by the generalizable and pattern-based technologies employed in different components of the framework. Another relevant feature of DemKG is the friendly interface of the transformation module, which lowers the technical barrier to effectively integrating study research data in the KG.

However, there exists an important limitation in its implementation: once built, the KG does not support modifications without risking underlying integrity, forcing a complete build and possibly ingestion when new versions become available. This limitation, a consequence of using KGX as the backbone for merging and building operations, may ultimately limit projects with streamed or on-demand data ingestion needs.

Nevertheless, our implementations remain open-source, primarily based on open knowledge sources, and the building pipelines employ systematic approaches with templating engines that are easily customizable. While our focus is dementia research, the broad biomedical ontologies forming the foundation of our terminological model make our KG applicable to other biomedical research datasets as well.

Thus, the broader implications of our work extend beyond the application of the KG. Large biomedical KGs are proving to be an excellent tool for biomedical research, especially in domains requiring knowledge across different fields. The capacity to integrate disparate data and knowledge opens up opportunities for insights that were previously challenging to achieve. Approaches such as Precision Medicine greatly benefit from the implementation of KGs in their workflow.

This benefit is especially pronounced in dementia research, where the number of newly discovered biomarkers, phenotypes, and life conditions rapidly increases. These elements become part of the knowledge base that can be applied to the patient's biological signature. In this context, a KG like ours can play a crucial role in advancing our understanding of dementia and potentially informing patient care strategies.

6. Conclusions

In conclusion, DemKG presents a flexible and integrative approach to handle the ever-increasing complexity and multi-modality of dementia research data by leveraging a KG representation and relation capabilities.

The DemKG framework offers several distinct advantages over other solutions currently available. First, it is constructed based on well-established ontologies and adheres to recognized community standards, guaranteeing a solid and interoperable foundation. This is further enhanced by ontological extensions specifically crafted to facilitate detailed dementia research data analysis, filling a critical gap in the existing frameworks.

In addition to the above, DemKG integrates a low-code transformer module, simplifying the integration of study data and making the framework accessible to researchers with various levels of expertise. This module significantly reduces the time and technical know-how needed to merge study data, streamlining the data integration process considerably when compared to other solutions.

Furthermore, DemKG employs tooling to generate knowledge graphs in the platform-agnostic KGX format. This approach allows for easy deployment in a platform of the user's choice, offering flexibility in how and where the data can be used, and ensuring that the framework is adaptable to existing systems and future technological advancements. Enhancing its flexibility, the framework offers an open-source and customizable design, facilitating easy adoption and adaptation not only for dementia research but also potentially extending its utility to research into other diseases.

While there are limitations to the support for post-build modifications in its current iteration, addressing these in future work could broaden its applicability further. Despite these challenges, DemKG and similar KGs hold significant potential for propelling biomedical research and patient care advancements, extending from dementia to other medical conditions.

Author Contributions: Conceptualization, S.T.-R., M.R. and R.M.-T.; methodology, S.T.-R., M.R. and R.M.-T.; software, S.T.-R.; validation, M.R., R.M.-T., B.-E.K. and T.F.; formal analysis, S.T.-R.; writing—original draft preparation, S.T.-R.; writing—review and editing, S.T.-R., M.R., R.M.-T., B.-E.K. and T.F.; visualization, S.T.-R.; supervision, M.R. and R.M.-T.; funding acquisition, T.F. and M.R. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Norwegian Research Council, JPND/PMI-AD (NRC 311993) and Dementia Disease Initiation (217780); Helse Sør-øst, NASATS Dementia Disease Initiation (2013131); Public-Private Partnership Research project CPP2021-009109 of the Spanish Program to Promote Scientific and Technological Research; and Research project PID2019-110686RB-I00 of the Spanish Research Program Oriented to the Challenges of Society.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data from the DDI study employed in the use case are not publicly available due to ethical and patient privacy restrictions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Manzoni, C.; Kia, D.A.; Vandrovcova, J.; Hardy, J.; Wood, N.W.; Lewis, P.A.; Ferrari, R. Genome, Transcriptome and Proteome: The Rise of Omics Data and Their Integration in Biomedical Sciences. *Brief. Bioinform.* **2018**, *19*, 286–302. [CrossRef] [PubMed]
2. Misra, B.B.; Langefeld, C.; Olivier, M.; Cox, L.A. Integrated Omics: Tools, Advances and Future Approaches. *J. Mol. Endocrinol.* **2019**, *62*, R21–R45. [CrossRef]
3. Glaab, E.; Rauschenberger, A.; Banzi, R.; Gerardi, C.; Garcia, P.; Demotes, J. Biomarker Discovery Studies for Patient Stratification Using Machine Learning Analysis of Omics Data: A Scoping Review. *BMJ Open* **2021**, *11*, e053674. [CrossRef] [PubMed]
4. Sun, Z.; Ng, K.; Ramli, N. Biomedical Imaging Research: A Fast-Emerging Area for Interdisciplinary Collaboration. *Biomed. Imaging Interv. J.* **2011**, *7*, e21. [CrossRef]
5. Lussier, Y.A.; Liu, Y. Computational Approaches to Phenotyping: High-Throughput Phenomics. *Proc. Am. Thorac. Soc.* **2007**, *4*, 18–25. [CrossRef] [PubMed]
6. Che, Z.; Liu, Y. Deep Learning Solutions to Computational Phenotyping in Health Care. In Proceedings of the IEEE International Conference on Data Mining Workshops (ICDMW), New Orleans, LA, USA, 18–21 November 2017; pp. 1100–1109.
7. Che, Z.; Kale, D.; Li, W.; Bahadori, M.T.; Liu, Y. Deep Computational Phenotyping. In Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, Australia, 10–13 August 2015; pp. 507–516.
8. Barabasi, A.-L.; Oltvai, Z.N. Network Biology: Understanding the Cell's Functional Organization. *Nat. Rev. Genet.* **2004**, *5*, 101–113. [CrossRef] [PubMed]
9. Barabási, A.-L.; Gulbahce, N.; Loscalzo, J. Network Medicine: A Network-Based Approach to Human Disease. *Nat. Rev. Genet.* **2011**, *12*, 56–68. [CrossRef]
10. Timón-Reina, S.; Rincón, M.; Martínez-Tomás, R. An Overview of Graph Databases and Their Applications in the Biomedical Domain. *Database* **2021**, *2021*, 26. [CrossRef]
11. Introducing the Knowledge Graph: Things, Not Strings. Available online: https://blog.google/products/search/introducing-knowledge-graph-things-not/ (accessed on 13 September 2023).
12. Noy, N.; Gao, Y.; Jain, A.; Narayanan, A.; Patterson, A.; Taylor, J. Industry-Scale Knowledge Graphs: Lessons and Challenges: Five Diverse Technology Companies Show How It's Done. *Queue* **2019**, *17*, 48–75. [CrossRef]
13. Sheth, A.; Padhee, S.; Gyrard, A. Knowledge Graphs and Knowledge Networks: The Story in Brief. *IEEE Internet Comput.* **2019**, *23*, 67–75. [CrossRef]
14. Ehrlinger, L.; Wöß, W. Towards a definition of knowledge graphs. In *CEUR Workshop Proceedings*; CEUR-WS: Aachen, Germany, 2016; Volume 1695.
15. Hogan, A.; Blomqvist, E.; Cochez, M.; d'Amato, C.; de Melo, G.; Gutierrez, C.; Kirrane, S.; Gayo, J.E.L.; Navigli, R.; Neumaier, S. Knowledge Graphs. *ACM Comput. Surv. CSUR* **2021**, *54*, 1–37.

16. Besta, M.; Peter, E.; Gerstenberger, R.; Fischer, M.; Podstawski, M.; Barthels, C.; Alonso, G.; Hoefler, T. Demystifying Graph Databases: Analysis and Taxonomy of Data Organization, System Designs, and Graph Queries. *ACM Comput. Surv.* **2019**, *56*, 1–40. [CrossRef]
17. Brandizi, M.; Singh, A.; Rawlings, C.; Hassani-Pak, K. Getting the Best of Linked Data and Property Graphs: Rdf2neo and the KnetMiner Use Case. In Proceedings of the CEUR Workshop Proceedings, Antwerp, Belgium, 3–6 December 2018; Volume 2275.
18. Alocci, D.; Mariethoz, J.; Horlacher, O.; Bolleman, J.T.; Campbell, M.P.; Lisacek, F. Property Graph vs RDF Triple Store: A Comparison on Glycan Substructure Search. *PLoS ONE* **2015**, *10*, e0144578. [CrossRef] [PubMed]
19. Hoehndorf, R.; Schofield, P.N.; Gkoutos, G.V. The Role of Ontologies in Biological and Biomedical Research: A Functional Perspective. *Brief. Bioinform.* **2015**, *16*, 1069–1080. [CrossRef]
20. Dovrolis, N.; Stefanut, T.; Dietze, S.; Yu, H.Q.; Valentine, C.; Kaldoudi, E. Semantic Annotation and Linking of Medical Educational Resources. In *5th European Conference of the International Federation for Medical and Biological Engineering 14–18 September 2011, Budapest, Hungary*; Jobbágy, Á., Ed.; IFMBE Proceedings; Springer: Berlin/Heidelberg, Germany, 2011; Volume 37, pp. 1400–1403.
21. Song, D.; Chute, C.G.; Tao, C. Semantator: Annotating Clinical Narratives with Semantic Web Ontologies. *AMIA Jt. Summits Transl. Sci. Proc.* **2012**, *2012*, 20–209. [PubMed]
22. Shah, N.H.; Bhatia, N.; Jonquet, C.; Rubin, D.; Chiang, A.P.; Musen, M.A. Comparison of concept recognizers for building the open biomedical annotator. In *BMC Bioinformatics*; BioMed Central: London, UK, 2009; Volume 10.
23. El-Haj, M.; Rutherford, N.; Coole, M.; Ezeani, I.; Prentice, S.; Ide, N.; Knight, J.; Piao, S.; Mariani, J.; Rayson, P.; et al. Infrastructure for Semantic Annotation in the Genomics Domain. In Proceedings of the LREC, Marseille, France, 20–25 June 2020.
24. Tan, H.; Lambrix, P. Selecting an ontology for biomedical text mining. In *Workshop on Current Trends in Biomedical Natural Language Processing*; Association for Computational Linguistics: Boulder, CO, USA, 2009; pp. 55–62.
25. Witte, R.; Kappler, T.; Baker, C.J.O. Ontology Design for Biomedical Text Mining. In *Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences*; Springer: Boston, MA, USA, 2007; Volume 9780387484, pp. 281–313. ISBN 978-0-387-48438-9.
26. Jackson, R.; Matentzoglu, N.; Overton, J.A.; Vita, R.; Balhoff, J.P.; Buttigieg, P.L.; Carbon, S.; Courtot, M.; Diehl, A.D.; Dooley, D.M.; et al. OBO Foundry in 2021: Operationalizing Open Data Principles to Evaluate Ontologies. *Database* **2021**, *2021*, baab069. [CrossRef] [PubMed]
27. Musen, M.A.; Noy, N.F.; Shah, N.H.; Whetzel, P.L.; Chute, C.G.; Story, M.A.; Smith, B. The National Center for Biomedical Ontology. *J. Am. Med. Inform. Assoc.* **2012**, *19*, 190–195. [CrossRef]
28. Whetzel, P.L.; Noy, N.F.; Shah, N.H.; Alexander, P.R.; Nyulas, C.; Tudorache, T.; Musen, M.A. BioPortal: Enhanced Functionality via New Web Services from the National Center for Biomedical Ontology to Access and Use Ontologies in Software Applications. *Nucleic Acids Res.* **2011**, *39*, W541–W545. [CrossRef] [PubMed]
29. Mungall, C.J.; McMurry, J.A.; Köhler, S.; Balhoff, J.P.; Borromeo, C.; Brush, M.; Carbon, S.; Conlin, T.; Dunn, N.; Engelstad, M.; et al. The Monarch Initiative: An Integrative Data and Analytic Platform Connecting Phenotypes to Genotypes across Species. *Nucleic Acids Res.* **2017**, *45*, D712–D722. [CrossRef] [PubMed]
30. Santos, A.; Colaço, A.R.; Nielsen, A.B.; Niu, L.; Strauss, M.; Geyer, P.E.; Coscia, F.; Albrechtsen, N.J.W.; Mundt, F.; Jensen, L.J.; et al. A Knowledge Graph to Interpret Clinical Proteomics Data. *Nat. Biotechnol.* **2022**, *40*, 692–702. [CrossRef] [PubMed]
31. Chandak, P.; Huang, K.; Zitnik, M. Building a Knowledge Graph to Enable Precision Medicine. *Sci. Data* **2023**, *10*, 67. [CrossRef] [PubMed]
32. Morris, J.H.; Soman, K.; Akbas, R.E.; Zhou, X.; Smith, B.; Meng, E.C.; Huang, C.C.; Cerono, G.; Schenk, G.; Rizk-Jackson, A. The Scalable Precision Medicine Open Knowledge Engine (SPOKE): A Massive Knowledge Graph of Biomedical Information. *Bioinformatics* **2023**, *39*, btad080. [CrossRef] [PubMed]
33. Reese, J.T.; Unni, D.; Callahan, T.J.; Cappelletti, L.; Ravanmehr, V.; Carbon, S.; Shefchek, K.A.; Good, B.M.; Balhoff, J.P.; Fontana, T.; et al. KG-COVID-19: A Framework to Produce Customized Knowledge Graphs for COVID-19 Response. *Patterns* **2021**, *2*, 100155. [CrossRef] [PubMed]
34. Badal, V.D.; Wright, D.; Katsis, Y.; Kim, H.-C.; Swafford, A.D.; Knight, R.; Hsu, C.-N. Challenges in the Construction of Knowledge Bases for Human Microbiome-Disease Associations. *Microbiome* **2019**, *7*, 129. [CrossRef] [PubMed]
35. Chaves-Fraga, D.; Endris, K.M.; Iglesias, E.; Corcho, O.; Vidal, M.-E. What Are the Parameters That Affect the Construction of a Knowledge Graph? In Proceedings of the On the Move to Meaningful Internet Systems: OTM 2019 Conferences: Confederated International Conferences: CoopIS, ODBASE, C&TC 2019, Rhodes, Greece, 21–25 October 2019; Springer: Berlin/Heidelberg, Germany, 2019; pp. 695–713.
36. Unni, D.R.; Moxon, S.A.T.; Bada, M.; Brush, M.; Bruskiewich, R.; Caufield, J.H.; Clemons, P.A.; Dancik, V.; Dumontier, M.; Fecho, K.; et al. Biolink Model: A Universal Schema for Knowledge Graphs in Clinical, Biomedical, and Translational Science. *Clin. Transl. Sci.* **2022**, *15*, 1848–1855. [CrossRef] [PubMed]
37. Caufield, J.H.; Putman, T.; Schaper, K.; Unni, D.R.; Hegde, H.; Callahan, T.J.; Cappelletti, L.; Moxon, S.A.; Ravanmehr, V.; Carbon, S.; et al. KG-Hub—Building and Exchanging Biological Knowledge Graphs 2023. *Bioinformatics* **2023**, *39*, btad418. [CrossRef] [PubMed]
38. Hamilton, W.L.; Ying, R.; Leskovec, J. Representation Learning on Graphs: Methods and Applications. *arXiv* **2017**, arXiv:1709.05584.
39. Chami, I.; Abu-El-Haija, S.; Perozzi, B.; Ré, C.; Murphy, K. Machine Learning on Graphs: A Model and Comprehensive Taxonomy. *J. Mach. Learn. Res.* **2022**, *23*, 3840–3903.

40. Cappelletti, L.; Fontana, T.; Casiraghi, E.; Ravanmehr, V.; Callahan, T.J.; Joachimiak, M.P.; Mungall, C.J.; Robinson, P.N.; Reese, J.; Valentini, G. GRAPE: Fast and Scalable Graph Processing and Embedding 2021. *Nat. Comput. Sci.* **2023**, *3*, 552–568. [CrossRef]
41. Ilievski, F.; Garijo, D.; Chalupsky, H.; Divvala, N.T.; Yao, Y.; Rogers, C.; Li, R.; Liu, J.; Singh, A.; Schwabe, D. KGTK: A Toolkit for Large Knowledge Graph Manipulation and Analysis. In Proceedings of the The Semantic Web–ISWC 2020: 19th International Semantic Web Conference, Athens, Greece, 2–6 November 2020; pp. 278–293.
42. Nelson, C.A.; Bove, R.; Butte, A.J.; Baranzini, S.E. Embedding Electronic Health Records onto a Knowledge Network Recognizes Prodromal Features of Multiple Sclerosis and Predicts Diagnosis. *J. Am. Med. Inform. Assoc.* **2022**, *29*, 424–434. [CrossRef] [PubMed]
43. Wang, Q.; Mao, Z.; Wang, B.; Guo, L. Knowledge Graph Embedding: A Survey of Approaches and Applications. *IEEE Trans. Knowl. Data Eng.* **2017**, *29*, 2724–2743. [CrossRef]
44. Li, X.; Chen, W.; Chen, Y.; Zhang, X.; Gu, J.; Zhang, M.Q. Network Embedding-Based Representation Learning for Single Cell RNA-Seq Data. *Nucleic Acids Res.* **2017**, *45*, e166. [CrossRef]
45. Liu, X.; Yang, Z.; Sang, S.; Lin, H.; Wang, J.; Xu, B. Detection of Protein Complexes from Multiple Protein Interaction Networks Using Graph Embedding. *Artif. Intell. Med.* **2019**, *96*, 107–115. [CrossRef]
46. Wang, X.; Gong, Y.; Yi, J.; Zhang, W. Predicting Gene-Disease Associations from the Heterogeneous Network Using Graph Embedding. In Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine—BIBM 2019, San Diego, CA, USA, 1 November 2019; pp. 504–511.
47. Xu, B.; Liu, Y.; Yu, S.; Wang, L.; Dong, J.; Lin, H.; Yang, Z.; Wang, J.; Xia, F. A Network Embedding Model for Pathogenic Genes Prediction by Multi-Path Random Walking on Heterogeneous Network. *BMC Med. Genom.* **2019**, *12*, 188. [CrossRef]
48. Malec, S.A.; Taneja, S.B.; Albert, S.M.; Elizabeth Shaaban, C.; Karim, H.T.; Levine, A.S.; Munro, P.; Callahan, T.J.; Boyce, R.D. Causal Feature Selection Using a Knowledge Graph Combining Structured Knowledge from the Biomedical Literature and Ontologies: A Use Case Studying Depression as a Risk Factor for Alzheimer's Disease. *J. Biomed. Inform.* **2023**, *142*, 104368. [CrossRef]
49. Nicholson, D.N.; Greene, C.S. Constructing Knowledge Graphs and Their Biomedical Applications. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 1414–1428. [CrossRef] [PubMed]
50. Arp, R.; Smith, B. Function, role and disposition in basic formal ontology. *Nat. Preced.* **2008**. [CrossRef]
51. Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T.; et al. Gene Ontology: Tool for The Unification of Biology. *Nat. Genet.* **2000**, *25*, 25–29. [CrossRef]
52. The Gene Ontology Consortium; Aleksander, S.A.; Balhoff, J.; Carbon, S.; Cherry, J.M.; Drabkin, H.J.; Ebert, D.; Feuermann, M.; Gaudet, P.; Harris, N.L.; et al. The Gene Ontology Knowledgebase in 2023. *Genetics* **2023**, *224*, iyad031. [CrossRef]
53. Hastings, J.; Owen, G.; Dekker, A.; Ennis, M.; Kale, N.; Muthukrishnan, V.; Turner, S.; Swainston, N.; Mendes, P.; Steinbeck, C. ChEBI in 2016: Improved Services and an Expanding Collection of Metabolites. *Nucleic Acids Res.* **2016**, *44*, D1214–D1219. [CrossRef] [PubMed]
54. Natale, D.A.; Arighi, C.N.; Barker, W.C.; Blake, J.A.; Bult, C.J.; Caudy, M.; Drabkin, H.J.; D'Eustachio, P.; Evsikov, A.V.; Huang, H. The Protein Ontology: A Structured Representation of Protein Forms and Complexes. *Nucleic Acids Res.* **2010**, *39*, D539–D545. [CrossRef]
55. Vasilevsky, N.A.; Matentzoglu, N.A.; Toro, S.; Flack, J.E.; Hegde, H.; Unni, D.R.; Alyea, G.F.; Amberger, J.S.; Babb, L.; Balhoff, J.P.; et al. Mondo: Unifying Diseases for the World, by the World. *medRxiv* **2022**. [CrossRef]
56. Köhler, S.; Doelken, S.C.; Mungall, C.J.; Bauer, S.; Firth, H.V.; Bailleul-Forestier, I.; Black, G.C.M.; Brown, D.L.; Brudno, M.; Campbell, J.; et al. The Human Phenotype Ontology Project: Linking Molecular Biology and Disease through Phenotype Data. *Nucleic Acids Res.* **2014**, *42*, D966–D974. [CrossRef] [PubMed]
57. Köhler, S.; Gargano, M.; Matentzoglu, N.; Carmody, L.C.; Lewis-Smith, D.; Vasilevsky, N.A.; Danis, D.; Balagura, G.; Baynam, G.; Brower, A.M.; et al. The Human Phenotype Ontology in 2021. *Nucleic Acids Res.* **2021**, *49*, D1207–D1217. [CrossRef]
58. Gkoutos, G.V.; Schofield, P.N.; Hoehndorf, R. The Anatomy of Phenotype Ontologies: Principles, Properties and Applications. *Brief. Bioinform.* **2018**, *19*, 1008–1021. [CrossRef] [PubMed]
59. Mungall, C.J.; Torniai, C.; Gkoutos, G.V.; Lewis, S.E.; Haendel, M.A. Uberon, an Integrative Multi-Species Anatomy Ontology. *Genome Biol.* **2012**, *13*, R5–R20. [CrossRef] [PubMed]
60. Haendel, M.A.; Balhoff, J.P.; Bastian, F.B.; Blackburn, D.C.; Blake, J.A.; Bradford, Y.; Comte, A.; Dahdul, W.M.; Dececchi, T.A.; Druzinsky, R.E. Unification of Multi-Species Vertebrate Anatomy Ontologies for Comparative Biology in Uberon. *J. Biomed. Semant.* **2014**, *5*, 21. [CrossRef]
61. Rosse, C.; Mejino, J.L.V. A Reference Ontology for Biomedical Informatics: The Foundational Model of Anatomy. *J. Biomed. Inform.* **2003**, *36*, 478–500. [CrossRef]
62. Cox, A.P.; Jensen, M.; Ruttenberg, A.; Szigeti, K.; Diehl, A.D. Measuring Cognitive Functions: Hurdles in the Development of the NeuroPsychological Testing Ontology. In Proceedings of the ICBO, Montreal, QC, Canada, 7–12 July 2013; pp. 78–83.
63. Gomez-Valades, A.; Martinez-Tomas, R.; Rincon, M. Integrative Base Ontology for the Research Analysis of Alzheimer's Disease-Related Mild Cognitive Impairment. *Front. Neuroinformatics* **2021**, *15*, 561691. [CrossRef]
64. Peters, B.; OBI Consortium, T. Ontology for Biomedical Investigations. *Nat. Preced.* **2009**, 1. [CrossRef]
65. Bandrowski, A.; Brinkman, R.; Brochhausen, M.; Brush, M.H.; Bug, B.; Chibucos, M.C.; Clancy, K.; Courtot, M.; Derom, D.; Dumontier, M. The Ontology for Biomedical Investigations. *PLoS ONE* **2016**, *11*, e0154556. [CrossRef]

66. Szklarczyk, D.; Kirsch, R.; Koutrouli, M.; Nastou, K.; Mehryary, F.; Hachilif, R.; Gable, A.L.; Fang, T.; Doncheva, N.T.; Pyysalo, S.; et al. The STRING Database in 2023: Protein–Protein Association Networks and Functional Enrichment Analyses for Any Sequenced Genome of Interest. *Nucleic Acids Res.* **2023**, *51*, D638–D646. [CrossRef]
67. Gillespie, M.; Jassal, B.; Stephan, R.; Milacic, M.; Rothfels, K.; Senff-Ribeiro, A.; Griss, J.; Sevilla, C.; Matthews, L.; Gong, C.; et al. The Reactome Pathway Knowledgebase 2022. *Nucleic Acids Res.* **2022**, *50*, D687–D692. [CrossRef] [PubMed]
68. Diehl, A.D.; Meehan, T.F.; Bradford, Y.M.; Brush, M.H.; Dahdul, W.M.; Dougall, D.S.; He, Y.; Osumi-Sutherland, D.; Ruttenberg, A.; Sarntivijai, S. The Cell Ontology 2016: Enhanced Content, Modularization, and Ontology Interoperability. *J. Biomed. Semant.* **2016**, *7*, 44. [CrossRef] [PubMed]
69. Nadendla, S.; Jackson, R.; Munro, J.; Quaglia, F.; Mészáros, B.; Olley, D.; Hobbs, E.T.; Goralski, S.M.; Chibucos, M.; Mungall, C.J.; et al. ECO: The Evidence and Conclusion Ontology, an Update for 2022. *Nucleic Acids Res.* **2022**, *50*, D1515–D1521. [CrossRef]
70. Malone, J.; Holloway, E.; Adamusiak, T.; Kapushesky, M.; Zheng, J.; Kolesnikov, N.; Zhukova, A.; Brazma, A.; Parkinson, H. Modeling Sample Variables with an Experimental Factor Ontology. *Bioinformatics* **2010**, *26*, 1112–1118. [CrossRef] [PubMed]
71. Mayer, G.; Montecchi-Palazzi, L.; Ovelleiro, D.; Jones, A.R.; Binz, P.-A.; Deutsch, E.W.; Chambers, M.; Kallhardt, M.; Levander, F.; Shofstahl, J.; et al. The HUPO Proteomics Standards Initiative- Mass Spectrometry Controlled Vocabulary. *Database* **2013**, *2013*, bat009. [CrossRef]
72. Stefancsik, R.; Balhoff, J.P.; Balk, M.A.; Ball, R.L.; Bello, S.M.; Caron, A.R.; Chesler, E.J.; de Souza, V.; Gehrke, S.; Haendel, M.; et al. The Ontology of Biological Attributes (OBA)—Computational Traits for the Life Sciences. *Mamm. Genome* **2023**, *34*, 364–378. [CrossRef]
73. Scheuermann, R.H.; Ceusters, W.; Smith, B. Toward an Ontological Treatment of Disease and Diagnosis. *Summit Transl. Bioinforma.* **2009**, *2009*, 116.
74. Hicks, A.; Hanna, J.; Welch, D.; Brochhausen, M.; Hogan, W.R. The Ontology of Medically Related Social Entities: Recent Developments. *J. Biomed. Semant.* **2016**, *7*, 47. [CrossRef] [PubMed]
75. Kurlowicz, L.; Wallace, M. The Mini-Mental State Examination (MMSE). *J. Gerontol. Nurs.* **1999**, *25*, 8–9. [CrossRef]
76. Fillenbaum, G.G.; Mohs, R. CERAD (Consortium to Establish a Registry for Alzheimer's Disease) Neuropsychology Assessment Battery: 35 Years and Counting. *J. Alzheimers Dis.* **2023**, *93*, 1–27. [CrossRef] [PubMed]
77. Quental, N.B.M.; Brucki, S.M.D.; Bueno, O.F.A. Visuospatial Function in Early Alzheimer's Disease—The Use of the Visual Object and Space Perception (VOSP) Battery. *PLoS ONE* **2013**, *8*, e68398. [CrossRef]
78. Bowie, C.R.; Harvey, P.D. Administration and Interpretation of the Trail Making Test. *Nat. Protoc.* **2006**, *1*, 2277–2281. [CrossRef]
79. Mainland, B.J.; Shulman, K.I. Clock Drawing Test. In *Cognitive Screening Instruments: A Practical Approach*; Springer: London, UK, 2017; pp. 67–108. [CrossRef]
80. Benton, A.L.; de Hamsher, S.; Sivan, A.B. Controlled Oral Word Association Test. *Arch. Clin. Neuropsychol.* **1994**. [CrossRef]
81. Jack, C.R., Jr.; Bennett, D.A.; Blennow, K.; Carrillo, M.C.; Dunn, B.; Haeberlein, S.B.; Holtzman, D.M.; Jagust, W.; Jessen, F.; Karlawish, J.; et al. NIA-AA Research Framework: Toward a Biological Definition of Alzheimer's Disease. *Alzheimers Dement.* **2018**, *14*, 535–562. [CrossRef] [PubMed]
82. Fecho, K.; Thessen, A.E.; Baranzini, S.E.; Bizon, C.; Hadlock, J.J.; Huang, S.; Roper, R.T.; Southall, N.; Ta, C.; Watkins, P.B.; et al. Progress toward a Universal Biomedical Data Translator. *Clin. Transl. Sci.* **2022**, *15*, 1838–1847. [CrossRef]
83. Matentzoglu, N.; Goutte-Gattat, D.; Tan, S.Z.K.; Balhoff, J.P.; Carbon, S.; Caron, A.R.; Duncan, W.D.; Flack, J.E.; Haendel, M.; Harris, N.L.; et al. Ontology Development Kit: A Toolkit for Building, Maintaining, and Standardising Biomedical Ontologies. *Database* **2022**, *2022*, baac087. [CrossRef]
84. Osumi-Sutherland, D.; Courtot, M.; Balhoff, J.P.; Mungall, C. Dead Simple OWL Design Patterns. *J. Biomed. Semant.* **2017**, *8*, 18. [CrossRef] [PubMed]
85. Hitzler, P.; Krötzsch, M.; Parsia, B.; Patel-Schneider, P.F.; Rudolph, S. OWL 2 Web Ontology Language. Available online: https://www.w3.org/TR/owl2-primer/ (accessed on 30 August 2023).
86. Lawrence Berkeley National Laboratory. (BBOP), Lawrence Berkeley National Knowledge Graph Hub. Available online: https://kghub.org/ (accessed on 30 August 2023).
87. KGX Format—Kgx 1.5.1 Documentation. Available online: https://kgx.readthedocs.io/en/latest/kgx_format.html (accessed on 30 August 2023).
88. KG-OBO. Available online: https://github.com/Knowledge-Graph-Hub/kg-obo (accessed on 30 August 2023).
89. Relation-Graph. Available online: https://github.com/INCATools/relation-graph (accessed on 30 August 2023).
90. Balhoff, J.P.; Bayindir, U.; Caron, A.R.; Matentzoglu, N.; Osumi-Sutherland, D.; Mungall, C.J. Ubergraph: Integrating OBO Ontologies into a Unified Semantic Graph. In *CEUR Workshop Proceedings*; CEUR-WS: Aachen, Germany, 2022; Volume 1613, p. 73.
91. Kostovska, A.; Tolovski, I.; Maikore, F.; Initiative, A.D.N.; Soldatova, L.; Panov, P. Neurodegenerative Disease Data Ontology. In Proceedings of the Discovery Science: 22nd International Conference, DS 2019, Split, Croatia, 28–30 October 2019; pp. 235–245.
92. Vita, R.; Zheng, J.; Jackson, R.; Dooley, D.; Overton, J.A.; Miller, M.A.; Berrios, D.C.; Scheuermann, R.H.; He, Y.; McGinty, H.K.; et al. Standardization of Assay Representation in the Ontology for Biomedical Investigations. *Database* **2021**, *2021*, baab040. [CrossRef]
93. Fischl, B. FreeSurfer. *Neuroimage* **2012**, *62*, 774–781. [CrossRef] [PubMed]

94. Yushkevich, P.A.; Pluta, J.B.; Wang, H.; Xie, L.; Ding, S.-L.; Gertje, E.C.; Mancuso, L.; Kliot, D.; Das, S.R.; Wolk, D.A. Automated Volumetry and Regional Thickness Analysis of Hippocampal Subfields and Medial Temporal Cortical Structures in Mild Cognitive Impairment. *Hum. Brain Mapp.* **2015**, *36*, 258–287. [CrossRef] [PubMed]
95. Basser, P.J.; Mattiello, J.; LeBihan, D. MR Diffusion Tensor Spectroscopy and Imaging. *Biophys. J.* **1994**, *66*, 259–267. [CrossRef] [PubMed]
96. Low, A.; Mak, E.; Stefaniak, J.D.; Malpetti, M.; Nicastro, N.; Savulich, G.; Chouliaras, L.; Markus, H.S.; Rowe, J.B.; O'Brien, J.T. Peak Width of Skeletonized Mean Diffusivity as a Marker of Diffuse Cerebrovascular Damage. *Front. Neurosci.* **2020**, *14*, 238. [CrossRef] [PubMed]
97. Fladby, T.; Pålhaugen, L.; Selnes, P.; Waterloo, K.; Bråthen, G.; Hessen, E.; Almdahl, I.S.; Arntzen, K.-A.; Auning, E.; Eliassen, C.F.; et al. Detecting At-Risk Alzheimer's Disease Cases. *J. Alzheimers Dis.* **2017**, *60*, 97–105. [CrossRef]
98. Marcus, D.S.; Olsen, T.R.; Ramaratnam, M.; Buckner, R.L. The Extensible Neuroimaging Archive Toolkit: An Informatics Platform for Managing, Exploring, and Sharing Neuroimaging Data. *Neuroinformatics* **2007**, *5*, 11–34. [CrossRef]
99. Fillenbaum, G.G.; van Belle, G.; Morris, J.C.; Mohs, R.C.; Mirra, S.S.; Davis, P.C.; Tariot, P.N.; Silverman, J.M.; Clark, C.M.; Welsh-Bohmer, K.A.; et al. Consortium to Establish a Registry for Alzheimer's Disease (CERAD): The First Twenty Years. *Alzheimers Dement.* **2008**, *4*, 96–109. [CrossRef] [PubMed]
100. Kirsebom, B.E.; Espenes, R.; Hessen, E.; Waterloo, K.; Johnsen, S.H.; Gundersen, E.; Botne Sando, S.; Rolfseng Grøntvedt, G.; Timón, S.; Fladby, T. Demographically Adjusted CERAD Wordlist Test Norms in a Norwegian Sample from 40 to 80 Years. *Clin. Neuropsychol.* **2019**, *33*, 27–39. [CrossRef]
101. Espenes, J.; Hessen, E.; Eliassen, I.V.; Waterloo, K.; Eckerström, M.; Sando, S.B.; Timón, S.; Wallin, A.; Fladby, T.; Kirsebom, B.-E. Demographically Adjusted Trail Making Test Norms in a Scandinavian Sample from 41 to 84 Years. *Clin. Neuropsychol.* **2020**, *34*, 110–126. [CrossRef]
102. Grover, A.; Leskovec, J. Node2vec: Scalable Feature Learning for Networks. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Association for Computing Machinery, San Francisco, CA, USA, 13–17 August 2016; pp. 855–864.
103. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv* **2013**, arXiv:1301.3781.
104. Van der Maaten, L.; Hinton, G. Visualizing Data Using T-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.

Chapter 6

Related publications

Apart from the main three articles, the development of this work has made different contributions to other published results, both directly and indirectly.

The first scientific contribution of this thesis was shared in the International Work-Conference on the Interplay Between Natural and Artificial Computation (IWINAC) with "*Towards an Integrated Semantic Framework for Neurological Multidimensional Data Analysis*" [Reina et al., 2015], where we outlined the first steps in developing the ISF.

As we described in the methods chapter, the collaboration with the DDI study has been fundamental for the development of both this thesis and the study. This work has partly contributed in the development of the research data ecosystem of DDI, firstly reported in "*Detecting At-Risk Alzheimer's Disease Cases*" [Fladby et al., 2017], by modeling the data structures to support the research outcomes of DDI, implement all related aspects in the XNAT platform, data capturing interfaces, and the KG-related improvements.

Finally, this work has contributed to the creation of regression-based norming from the DDI cohort. It has aided in data processing for normative data in various neuropsychological tests and has supported the development of XNAT tooling and open-source norm calculators for the scientific community. The resulting publications are:

- "*Demographically adjusted CERAD wordlist test norms in a Norwegian sample from 40 to 80 years*" [Kirsebom et al., 2019].
- "*Demographically adjusted trail making test norms in a Scandinavian sample from 41 to 84 years*" [Espenes et al., 2020].
- "*Regression-based norms for the FAS phonemic fluency test for ages 40–84 based on a Norwegian sample*" [Lorentzen et al., 2021].
- "*Regression-based cognitive change norms applied in biochemically defined predementia Alzheimer's disease*" [Eliassen et al., 2022].

- "*Regression-based normative data for the Rey Auditory Verbal Learning Test in Norwegian and Swedish adults aged 49–79 and comparison with published norms*" [Espenes et al., 2022].

Chapter 7

Conclusions

With the proliferation of data in biomedical research due to technological advancements, the challenge of effectively interconnecting diverse data modalities—such as omics data, imaging studies, and clinical records—has become increasingly pressing. KGs offer a robust framework for navigating and linking concepts across these varied data types and levels of abstraction, both in biological and mathematical contexts. By serving as an integrative layer, KGs facilitate the consolidation of information from disparate biological databases, including genomics, proteomics, and radiological imaging repositories. This aggregation of essential knowledge, which is often scattered across isolated platforms, has far-reaching implications for the concept of Precision Medicine. The capacity to seamlessly link data and concepts, from molecular and omics analyses to imaging studies and up to disease mechanisms and phenotypic descriptions, is of great importance. This integrative approach is especially relevant in the current landscape of Dementia research, where extensive efforts are being made to profile the biological underpinnings of disease mechanisms in an inherently complex and nuanced reality.

The overarching aim of this thesis was to explore and advance the use of semantic technologies, graph databases, and knowledge graphs in the biomedical domain, with a particular focus on Dementia research, ultimately producing DemKG. This open-source framework brings the building of KGs closer to the research groups.

The contributions of this work are multiple and significant. The first article laid out the foundational design principles of a framework that enhances neuroimaging and biobanking systems through the integration of semantic technologies. Designed to be modular and scalable, the framework integrates smoothly with existing systems such as XNAT. It adopts a layered architecture incorporating schemas, ontologies, and services, enabling semantic data access. This architecture is particularly effective in managing data across various levels of abstraction, ranging from raw data to more refined logical concepts. The use cases demonstrated the efficacy of this approach, particularly in the neuroscience domain,

where multidimensional datasets are common, confirming the framework's efficacy in such a context.

In light of the growth of GDBMSs, the second article provided a comprehensive review of GDBMSs and their applicability in the biomedical domain. While Relational Database Management Systems (RDBMSs) and other NoSQL engines are the go-to choice in many application settings, GDBMSs excel in handling densely connected datasets with complex relationships. The article highlighted four key advantages of GDBMSs: natural modeling of many-to-many relationships, intuitive query languages, schema flexibility, and superior performance in relationship-centric searches. These features are particularly beneficial in biomedical research, where data integration and complex network analyses are common. Given the evidence supporting the suitability of GDBMSs for biomedical data and the variety of available platforms, we designed DemKG to construct platform-agnostic KGs, thereby leaving the choice of platform to the ultimate adopters.

The third article formally introduced DemKG, laying out the design details and implementation to manage complex and multimodal data in Dementia research. In contrast with related works, DemKG offers several advantages, including its foundation on well-established ontologies, a low-code transformer module for easy data integration, and the mentioned platform-agnostic design. The framework fills a critical gap in existing solutions by providing detailed Dementia research data analysis capabilities. Despite some limitations in support for post-build modifications, DemKG holds significant potential for advancing biomedical research and patient care, not just in Dementia but also in other medical conditions.

The work presented herein demonstrates the utility and adaptability of these technologies in addressing complex, multimodal challenges in biomedical research, showcased by several real-world scenarios faced in collaboration with the DDI and PMI-AD studies.

Future work

For the next phase of research, there are several avenues for future investigation and development, aiming to address existing limitations and explore new functionalities and applications of the framework.

Regarding the framework implementation, there are various interventions planned. The initial focus will be on enhancing the support for post-build modifications in DemKG, particularly when deployed on Graph Database Management Systems like Neo4j. This issue will necessitate contributions to the KGX codebase to effectively manage conflicts within node categories across different knowledge graph builds. Another planned improvement is the incorporation of a YAML schema validator within the dataset descriptor of the transformer

module. This addition aims to identify and alert users to any errors or inconsistencies in the provided input descriptor.

As Graph-based Machine Learning (ML) techniques gain traction across various research and industrial domains, many GDBMS platforms, including Neo4j, are evolving to incorporate more than just traditional graph metrics and algorithms. They are now integrating representation techniques such as embeddings and linked vectors directly into their systems, obviating the need for external computations or the maintenance of parallel data structures. In light of this, we aim to extend DemKG to better integrate with these emerging graph-ML techniques. One first approach is including the Network Embedding All the Things (NEAT)[1] pipeline engine within the KG builder module of the framework.

Related to these integration improvements, we plan to exploit DemKG in several ongoing investigations about pre-dementia. The first one will consist of further improving the results of synaptic-related proteomics by exploring available information about proteins of interest and evaluate, in a systematic way, their role in a subset of the genomic sub-network. For another application, we plan to compute fine-tuned graph embeddings to feed several deep-learning architectures to improve prediction models further for amyloid pathology. One last intervention will address the integration of drug information from DrugBank [Wishart et al., 2018] to assist in drug repurposing investigations for several AD-related target molecules, using both knowledge exploration and graph embeddings.

Finally, we plan to expand the terminological scope of DemKG to improve the ontological descriptions of various biomarkers associated with Parkinson's Disease. Additionally, we intend to develop axioms and annotations to assert explicitly shared mechanisms across different neurodegenerative diseases in the model.

[1] NEAT: https://github.com/Knowledge-Graph-Hub/neat-ml#network-embedding-all-the-things-neat

Bibliography

Claudia Manzoni, Demis A Kia, Jana Vandrovcova, John Hardy, Nicholas W Wood, Patrick A Lewis, and Raffaele Ferrari. Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. *Briefings in Bioinformatics*, 19(2):286–302, March 2018. ISSN 1477-4054. doi: 10.1093/bib/bbw114. URL https://doi.org/10.1093/bib/bbw114.

Biswapriya B. Misra, Carl Langefeld, Michael Olivier, and Laura A. Cox. Integrated omics: tools, advances and future approaches. *Journal of Molecular Endocrinology*, 62(1):R21–R45, January 2019. ISSN 1479-6813, 0952-5041. doi: 10.1530/JME-18-0055. URL https://jme.bioscientifica.com/view/journals/jme/62/1/JME-18-0055.xml. Publisher: Bioscientifica Ltd Section: Journal of Molecular Endocrinology.

Enrico Glaab, Armin Rauschenberger, Rita Banzi, Chiara Gerardi, Paula Garcia, and Jacques Demotes. Biomarker discovery studies for patient stratification using machine learning analysis of omics data: a scoping review. *BMJ Open*, 11(12):e053674, December 2021. ISSN 2044-6055, 2044-6055. doi: 10.1136/bmjopen-2021-053674. URL https://bmjopen.bmj.com/content/11/12/e053674. Publisher: British Medical Journal Publishing Group Section: Patient-centred medicine.

Z Sun, KH Ng, and N Ramli. Biomedical imaging research: a fast-emerging area for interdisciplinary collaboration. *Biomedical Imaging and Intervention Journal*, 7(3):e21, July 2011. ISSN 1823-5530. doi: 10.2349/biij.7.3.e21. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3265193/.

Yves A. Lussier and Yang Liu. Computational approaches to phenotyping: high-throughput phenomics. *Proceedings of the American Thoracic Society*, 4(1):18–25, 2007. ISBN: 1546-3222 Publisher: American Thoracic Society.

Zhengping Che and Yan Liu. Deep learning solutions to computational phenotyping in health care. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 1100–1109. IEEE, 2017. ISBN 1-5386-3800-2.

Zhengping Che, David Kale, Wenzhe Li, Mohammad Taha Bahadori, and Yan Liu. Deep computational phenotyping. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 507–516, 2015.

Jorge L. Del-Aguila, Maria Victoria Fernández, Suzanne Schindler, Laura Ibanez, Yuetiva Deming, Shengmei Ma, Ben Saef, Kathleen Black, John Budde, Joanne Norton, Rachel Chasse, Alzheimer's Disease Neuroimaging Initiative (ADNI), Oscar Harari, Alison Goate, Chengjie Xiong, John C. Morris, and Carlos Cruchaga. Assessment of the Genetic Architecture of Alzheimer's Disease Risk in Rate of Memory Decline. *Journal of Alzheimer's Disease*, 62(2):745–756, 2018. ISSN 1875-8908. doi: 10.3233/JAD-170834. Publisher: IOS Press.

Ganna Leonenko, Maryam Shoai, Eftychia Bellou, Rebecca Sims, Julie Williams, John Hardy, Valentina Escott-Price, and the Alzheimer's Disease Neuroimaging Initiative. Genetic risk for alzheimer disease is distinct from genetic risk for amyloid deposition. *Annals of Neurology*, 86(3):427–435, 2019. ISSN 1531-8249. doi: 10.1002/ana.25530. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/ana.25530. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ana.25530.

Byron Creese, Ryan Arathimos, Helen Brooker, Dag Aarsland, Anne Corbett, Cathryn Lewis, Clive Ballard, and Zahinoor Ismail. Genetic risk for Alzheimer's disease, cognition, and mild behavioral impairment in healthy older adults. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 13(1):e12164, January 2021. ISSN 2352-8729. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/dad2.12164. Publisher: John Wiley & Sons, Ltd.

Jacob Brain, Leanne Greene, Eugene Y. H. Tang, Jennie Louise, Amy Salter, Sarah Beach, Deborah Turnbull, Mario Siervo, Blossom C. M. Stephan, and Phillip J. Tully. Cardiovascular disease, associated risk factors, and risk of dementia: An umbrella review of meta-analyses. *Frontiers in Epidemiology*, 3, 2023. ISSN 2674-1199. URL https://www.frontiersin.org/articles/10.3389/fepid.2023.1095236.

Javier María Peralta Ramos, Denise Kviatcovsky, and Michal Schwartz. Targeting the immune system towards novel therapeutic avenues to fight brain aging and neurodegeneration. *European Journal of Neuroscience*, 56(9):5413–5427, 2022. ISSN 1460-9568. doi: 10.1111/ejn.15609. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ejn.15609. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ejn.15609.

Róisín M. McManus. The Role of Immunity in Alzheimer's Disease. *Advanced Biology*, 6(5):2101166, 2022. ISSN 2701-0198. doi: 10.1002/adbi.202101166. URL

https://onlinelibrary.wiley.com/doi/abs/10.1002/adbi.202101166. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/adbi.202101166.

Lewis O. J. Killin, John M. Starr, Ivy J. Shiue, and Tom C. Russ. Environmental risk factors for dementia: a systematic review. *BMC Geriatrics*, 16(1):175, October 2016. ISSN 1471-2318. doi: 10.1186/s12877-016-0342-y. URL https://doi.org/10.1186/s12877-016-0342-y.

Yong-Li Zhao, Yi Qu, Ya-Nan Ou, Ya-Ru Zhang, Lan Tan, and Jin-Tai Yu. Environmental factors and risks of cognitive impairment and dementia: A systematic review and meta-analysis. *Ageing Research Reviews*, 72:101504, December 2021. ISSN 1568-1637. doi: 10.1016/j.arr.2021.101504. URL https://www.sciencedirect.com/science/article/pii/S1568163721002518.

Massimiliano Izzo. *Biomedical Research and Integrated Biobanking: An Innovative Paradigm for Heterogeneous Data Management*. 2016. ISBN 978-3-642-35132-7. doi: 10.1007/978-3-642-35133-4. arXiv: 1106.3562 ISSN: 3642351336.

Paul A. Harris, Robert Taylor, Brenda L. Minor, Veida Elliott, Michelle Fernandez, Lindsay O'Neal, Laura McLeod, Giovanni Delacqua, Francesco Delacqua, and Jacqueline Kirby. The REDCap consortium: building an international community of software platform partners. *Journal of biomedical informatics*, 95:103208, 2019. ISBN: 1532-0464 Publisher: Elsevier.

Daniel S Marcus, Timothy R Olsen, Mohana Ramaratnam, and Randy L Buckner. The Extensible Neuroimaging Archive Toolkit: an informatics platform for managing, exploring, and sharing neuroimaging data. *Neuroinformatics*, 5(1):11–34, January 2007. ISSN 1539-2791. URL http://www.ncbi.nlm.nih.gov/pubmed/17426351.

Adam Scott, Will Courtney, Dylan Wood, Raul de la Garza, Susan Lane, Margaret King, Runtang Wang, Jody Roberts, Jessica A. Turner, and Vince D. Calhoun. COINS: An Innovative Informatics and Neuroimaging Tool Suite Built for Large Heterogeneous Datasets. *Frontiers in Neuroinformatics*, 5:33, 2011. ISSN 1662-5196. doi: 10.3389/fninf.2011.00033. URL http://journal.frontiersin.org/article/10.3389/fninf.2011.00033/abstract. Publisher: Frontiers.

Luca Corradi, Gabriele Arnulfo, Andrea Schenone, Ivan Porro, and Marco Fato. XTENS - an eXTensible environment for neuroscience. *Studies in health technology and informatics*, 147:127–36, 2009. ISSN 0926-9630. URL http://www.ncbi.nlm.nih.gov/pubmed/19593051.

Rebecca Jackson, Nicolas Matentzoglu, James A Overton, Randi Vita, James P Balhoff, Pier Luigi Buttigieg, Seth Carbon, Melanie Courtot, Alexander D Diehl, Damion M Dooley, William D Duncan, Nomi L Harris, Melissa A Haendel, Suzanna E Lewis, Darren A Natale, David Osumi-Sutherland, Alan Ruttenberg, Lynn M Schriml, Barry Smith, Christian J Stoeckert Jr., Nicole A Vasilevsky, Ramona L Walls, Jie Zheng, Christopher J Mungall, and Bjoern Peters. OBO Foundry in 2021: operationalizing open data principles to evaluate ontologies. *Database*, 2021:baab069, September 2021. ISSN 1758-0463. doi: 10.1093/database/baab069. URL https://doi.org/10.1093/database/baab069.

Mark A. Musen, Natalya F. Noy, Nigam H. Shah, Patricia L. Whetzel, Christopher G. Chute, Margaret Anne Story, and Barry Smith. The National center for biomedical ontology. *Journal of the American Medical Informatics Association*, 19(2):190–195, March 2012. ISSN 10675027. doi: 10.1136/amiajnl-2011-000523. URL https://academic.oup.com/jamia/article-lookup/doi/10.1136/amiajnl-2011-000523. Publisher: Oxford Academic.

Patricia L Whetzel, Natalya F Noy, Nigam H Shah, Paul R Alexander, Csongor Nyulas, Tania Tudorache, and Mark A Musen. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic acids research*, 39(Web Server issue):W541–5, July 2011. ISSN 1362-4962. doi: 10.1093/nar/gkr469. URL http://nar.oxfordjournals.org/content/39/suppl_2/W541.short.

Béla Bollobás. *Modern Graph Theory*, volume 184. Springer-Verlag New York, 1998. ISBN 978-0-387-98488-9. doi: 10.1007/978-1-4612-0619-4_1. URL http://link.springer.com/10.1007/978-1-4612-0619-4_1.

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, and Sebastian Neumaier. Knowledge graphs. *ACM Computing Surveys (CSUR)*, 54(4):1–37, 2021. ISBN: 0360-0300 Publisher: ACM New York, NY, USA.

Amit Sheth, Swati Padhee, and Amelie Gyrard. Knowledge Graphs and Knowledge Networks: The Story in Brief. *IEEE Internet Computing*, 23(4):67–75, July 2019. ISSN 19410131. doi: 10.1109/MIC.2019.2928449. Publisher: Institute of Electrical and Electronics Engineers Inc.

Lisa Ehrlinger and Wolfram Wöß. Towards a Definition of Knowledge Graphs. *SEMANTiCS (Posters, Demos, SuCCESS)*, 48, 2016.

Natasha Noy, Yuqing Gao, Anshu Jain, Anant Narayanan, Alan Patterson, and Jamie Taylor. Industry-scale Knowledge Graphs: Lessons and Challenges: Five diverse technology companies show how it's done. *Queue*, 17(2):48–75, 2019. ISBN: 1542-7730 Publisher: ACM New York, NY, USA.

Albert-Laszlo Barabasi and Zoltan N. Oltvai. Network biology: understanding the cell's functional organization. *Nature reviews genetics*, 5(2):101–113, 2004. ISBN: 1471-0056 Publisher: Nature Publishing Group UK London.

Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. Network medicine: a network-based approach to human disease. *Nature reviews genetics*, 12(1):56–68, 2011. ISBN: 1471-0056 Publisher: Nature Publishing Group UK London.

David N. Nicholson and Casey S. Greene. Constructing knowledge graphs and their biomedical applications. *Computational and Structural Biotechnology Journal*, 18: 1414–1428, January 2020. ISSN 20010370. doi: 10.1016/j.csbj.2020.05.017. URL https://linkinghub.elsevier.com/retrieve/pii/S2001037020302804. Publisher: Elsevier B.V.

Christopher J. Mungall, Julie A. McMurry, Sebastian Köhler, James P. Balhoff, Charles Borromeo, Matthew Brush, Seth Carbon, Tom Conlin, Nathan Dunn, Mark Engelstad, Erin Foster, J.P. Gourdine, Julius O.B. Jacobsen, Dan Keith, Bryan Laraway, Suzanna E. Lewis, Jeremy NguyenXuan, Kent Shefchek, Nicole Vasilevsky, Zhou Yuan, Nicole Washington, Harry Hochheiser, Tudor Groza, Damian Smedley, Peter N. Robinson, and Melissa A. Haendel. The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Research*, 45(D1):D712–D722, January 2017. ISSN 0305-1048. doi: 10.1093/nar/gkw1128. URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw1128. Publisher: Narnia.

Alberto Santos, Ana R. Colaço, Annelaura B. Nielsen, Lili Niu, Maximilian Strauss, Philipp E. Geyer, Fabian Coscia, Nicolai J. Wewer Albrechtsen, Filip Mundt, Lars Juhl Jensen, and Matthias Mann. A knowledge graph to interpret clinical proteomics data. *Nature Biotechnology*, 40(5):692–702, May 2022. ISSN 1546-1696. doi: 10.1038/s41587-021-01145-6. URL https://www.nature.com/articles/s41587-021-01145-6. Number: 5 Publisher: Nature Publishing Group.

Payal Chandak, Kexin Huang, and Marinka Zitnik. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10(1):67, February 2023. ISSN 2052-4463. doi: 10.1038/s41597-023-01960-3. URL https://www.nature.com/articles/s41597-023-01960-3. Number: 1 Publisher: Nature Publishing Group.

John H. Morris, Karthik Soman, Rabia E. Akbas, Xiaoyuan Zhou, Brett Smith, Elaine C. Meng, Conrad C. Huang, Gabriel Cerono, Gundolf Schenk, and Angela Rizk-Jackson. The scalable precision medicine open knowledge engine (SPOKE): a massive knowledge graph of biomedical information. *Bioinformatics*, 39(2):btad080, 2023. doi: 10.1093/bioinformatics/btad080. URL https://academic.oup.com/bioinformatics/article/39/2/btad080/7033465. ISBN: 1367-4811 Publisher: Oxford University Press.

Varsha Dave Badal, Dustin Wright, Yannis Katsis, Ho-Cheol Kim, Austin D. Swafford, Rob Knight, and Chun-Nan Hsu. Challenges in the construction of knowledge bases for human microbiome-disease associations. *Microbiome*, 7(1):129, September 2019. ISSN 2049-2618. doi: 10.1186/s40168-019-0742-2. URL https://doi.org/10.1186/s40168-019-0742-2.

David Chaves-Fraga, Kemele M. Endris, Enrique Iglesias, Oscar Corcho, and Maria-Esther Vidal. What are the parameters that affect the construction of a knowledge graph? In *On the Move to Meaningful Internet Systems: OTM 2019 Conferences: Confederated International Conferences: CoopIS, ODBASE, C&TC 2019, Rhodes, Greece, October 21–25, 2019, Proceedings*, pages 695–713. Springer, 2019. ISBN 3-030-33245-4.

Deepak R. Unni, Sierra A. T. Moxon, Michael Bada, Matthew Brush, Richard Bruskiewich, J. Harry Caufield, Paul A. Clemons, Vlado Dancik, Michel Dumontier, Karamarie Fecho, Gustavo Glusman, Jennifer J. Hadlock, Nomi L. Harris, Arpita Joshi, Tim Putman, Guangrong Qin, Stephen A. Ramsey, Kent A. Shefchek, Harold Solbrig, Karthik Soman, Anne E. Thessen, Melissa A. Haendel, Chris Bizon, Christopher J. Mungall, and The Biomedical Data Translator Consortium. Biolink Model: A universal schema for knowledge graphs in clinical, biomedical, and translational science. *Clinical and Translational Science*, 15(8):1848–1855, 2022. ISSN 1752-8062. doi: 10.1111/cts.13302. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/cts.13302. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cts.13302.

J. Harry Caufield, Tim Putman, Kevin Schaper, Deepak R. Unni, Harshad Hegde, Tiffany J. Callahan, Luca Cappelletti, Sierra AT Moxon, Vida Ravanmehr, Seth Carbon, Lauren E. Chan, Katherina Cortes, Kent A. Shefchek, Glass Elsarboukh, James P. Balhoff, Tommaso Fontana, Nicolas Matentzoglu, Richard M. Bruskiewich, Anne E. Thessen, Nomi L. Harris, Monica C. Munoz-Torres, Melissa A. Haendel, Peter N. Robinson, Marcin P. Joachimiak, Christopher J. Mungall, and Justin T. Reese. KG-Hub – Building and Exchanging Biological Knowledge Graphs, January 2023. URL http://arxiv.org/abs/2302.10800. arXiv:2302.10800 [cs, q-bio].

Nicole A. Vasilevsky, Nicolas A. Matentzoglu, Sabrina Toro, Joseph E. Flack, Harshad Hegde, Deepak R. Unni, Gioconda F. Alyea, Joanna S. Amberger, Larry Babb, James P. Balhoff, Taylor I. Bingaman, Gully A. Burns, Orion J. Buske, Tiffany J. Callahan, Leigh C. Carmody, Paula Carrio Cordo, Lauren E. Chan, George S. Chang, Sean L. Christiaens, Louise C. Daugherty, Michel Dumontier, Laura E. Failla, May J. Flowers, H. Alpha Garrett, Jennifer L. Goldstein, Dylan Gration, Tudor Groza, Marc Hanauer, Nomi L. Harris, Jason A. Hilton, Daniel S. Himmelstein, Charles Tapley Hoyt, Megan S. Kane, Sebastian Köhler, David Lagorce, Abbe Lai, Martin Larralde, Antonia Lock, Irene López Santiago, Donna R. Maglott, Adriana J. Malheiro, Birgit H. M. Meldal, Monica C. Munoz-Torres, Tristan H. Nelson, Frank W. Nicholas, David Ochoa, Daniel P. Olson, Tudor I. Oprea, David Osumi-Sutherland, Helen Parkinson, Zoë May Pendlington, Ana Rath, Heidi L. Rehm, Lyubov Remennik, Erin R. Riggs, Paola Roncaglia, Justyne E. Ross, Marion F. Shadbolt, Kent A. Shefchek, Morgan N. Similuk, Nicholas Sioutos, Damian Smedley, Rachel Sparks, Ray Stefancsik, Ralf Stephan, Andrea L. Storm, Doron Stupp, Gregory S. Stupp, Jagadish Chandrabose Sundaramurthi, Imke Tammen, Darin Tay, Courtney L. Thaxton, Eloise Valasek, Jordi Valls-Margarit, Alex H. Wagner, Danielle Welter, Patricia L. Whetzel, Lori L. Whiteman, Valerie Wood, Colleen H. Xu, Andreas Zankl, Xingmin Aaron Zhang, Christopher G. Chute, Peter N. Robinson, Christopher J. Mungall, Ada Hamosh, and Melissa A. Haendel. Mondo: Unifying diseases for the world, by the world, May 2022. URL https://www.medrxiv.org/content/10.1101/2022.04.13.22273750v3. Pages: 2022.04.13.22273750.

S. Timón, M. Rincón, and R. Martínez-Tomás. Extending XNAT platform with an incremental semantic framework. *Frontiers in Neuroinformatics*, 11, 2017. ISSN 16625196. doi: 10.3389/fninf.2017.00057.

Santiago Timón-Reina, Mariano Rincón, and Rafael Martínez-Tomás. An overview of graph databases and their applications in the biomedical domain. *Database*, 2021: 26, October 2021. doi: 10.1093/DATABASE/BAAB026. URL https://academic.oup.com/database/article/doi/10.1093/database/baab026/6277712. Publisher: Oxford Academic.

Santiago Timón-Reina, Mariano Rincón, Rafael Martínez-Tomás, Bjørn-Eivind Kirsebom, and Tormod Fladby. A Knowledge Graph Framework for Dementia Research Data. *Applied Sciences*, 13(18):10497, January 2023. ISSN 2076-3417. doi: 10.3390/app131810497. URL https://www.mdpi.com/2076-3417/13/18/10497. Number: 18 Publisher: Multidisciplinary Digital Publishing Institute.

T. Fladby, L. Palhaugen, P. Selnes, K. Waterloo, G. Brathen, E. Hessen, I.S. Almdahl, K.-A. Arntzen, E. Auning, C.F. Eliassen, R. Espenes, R. Grambaite, G.R. Grøntvedt,

K.K. Johansen, S.H. Johnsen, L.F. Kalheim, B.-E. Kirsebom, K.I. Muller, A.E. Nakling, A. Rongven, S.B. Sando, N. Siafarikas, A.L. Stav, S. Tecelao, S. Timon, S.I. Bekkelund, and D. Aarsland. Detecting at-risk Alzheimer's disease cases. *Journal of Alzheimer's Disease*, 60(1), 2017. ISSN 18758908. doi: 10.3233/JAD-170231.

Alex J. Mitchell, Vicky Bird, Maria Rizzo, and Nick Meader. Diagnostic validity and added value of the geriatric depression scale for depression in primary care: A meta-analysis of GDS30 and GDS15. *Journal of Affective Disorders*, 125(1-3):10–17, September 2010. ISSN 01650327. doi: 10.1016/j.jad.2009.08.019. URL http://www.ncbi.nlm.nih.gov/pubmed/19800132.

Lenore Kurlowicz and Meredith Wallace. The mini-mental state examination (MMSE), 1999. Issue: 5 Pages: 8–9 Publication Title: Journal of gerontological nursing Volume: 25.

Brian J. Mainland and Kenneth I. Shulman. Clock drawing test. *Cognitive screening instruments: A practical approach*, pages 67–108, 2017. doi: https://doi.org/10.1007/978-3-319-44775-9_5. ISBN: 3319447742 Publisher: Springer.

Gerda G. Fillenbaum and Richard Mohs. CERAD (Consortium to Establish a Registry for Alzheimer's Disease) Neuropsychology Assessment Battery: 35 Years and Counting. *Journal of Alzheimer's Disease*, (Preprint):1–27, 2023. ISBN: 1387-2877 Publisher: IOS Press.

Natália Bezerra Mota Quental, Sonia Maria Dozzi Brucki, and Orlando Francisco Amodeo Bueno. Visuospatial function in early Alzheimer's disease—the use of the Visual Object and Space Perception (VOSP) battery. *PloS one*, 8(7):e68398, 2013. Publisher: Public Library of Science San Francisco, USA.

Christopher R. Bowie and Philip D. Harvey. Administration and interpretation of the Trail Making Test. *Nature protocols*, 1(5):2277–2281, 2006. Publisher: Nature Publishing Group UK London.

A. L. Benton, de SK Hamsher, and A. B. Sivan. Controlled oral word association test. *Archives of Clinical Neuropsychology*, 1994. doi: https://doi.org/10.1037/t10132-000.

Karamarie Fecho, Anne E. Thessen, Sergio E. Baranzini, Chris Bizon, Jennifer J. Hadlock, Sui Huang, Ryan T. Roper, Noel Southall, Casey Ta, Paul B. Watkins, Mark D. Williams, Hao Xu, William Byrd, Vlado Dančík, Marc P. Duby, Michel Dumontier, Gustavo Glusman, Nomi L. Harris, Eugene W. Hinderer, Greg Hyde, Adam Johs, Andrew I. Su, Guangrong Qin, Qian Zhu, and The Biomedical Data Translator Consortium. Progress toward a universal biomedical data translator. *Clinical and*

Translational Science, 15(8):1838–1847, 2022. ISSN 1752-8062. doi: 10.1111/cts.13301. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/cts.13301. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cts.13301.

The Gene Ontology Consortium, Suzi A Aleksander, James Balhoff, Seth Carbon, J Michael Cherry, Harold J Drabkin, Dustin Ebert, Marc Feuermann, Pascale Gaudet, Nomi L Harris, David P Hill, Raymond Lee, Huaiyu Mi, Sierra Moxon, Christopher J Mungall, Anushya Muruganugan, Tremayne Mushayahama, Paul W Sternberg, Paul D Thomas, Kimberly Van Auken, Jolene Ramsey, Deborah A Siegele, Rex L Chisholm, Petra Fey, Maria Cristina Aspromonte, Maria Victoria Nugnes, Federica Quaglia, Silvio Tosatto, Michelle Giglio, Suvarna Nadendla, Giulia Antonazzo, Helen Attrill, Gil dos Santos, Steven Marygold, Victor Strelets, Christopher J Tabone, Jim Thurmond, Pinglei Zhou, Saadullah H Ahmed, Praoparn Asanitthong, Diana Luna Buitrago, Meltem N Erdol, Matthew C Gage, Mohamed Ali Kadhum, Kan Yan Chloe Li, Miao Long, Aleksandra Michalak, Angeline Pesala, Armalya Pritazahra, Shirin C C Saverimuttu, Renzhi Su, Kate E Thurlow, Ruth C Lovering, Colin Logie, Snezhana Oliferenko, Judith Blake, Karen Christie, Lori Corbani, Mary E Dolan, Harold J Drabkin, David P Hill, Li Ni, Dmitry Sitnikov, Cynthia Smith, Alayne Cuzick, James Seager, Laurel Cooper, Justin Elser, Pankaj Jaiswal, Parul Gupta, Pankaj Jaiswal, Sushma Naithani, Manuel Lera-Ramirez, Kim Rutherford, Valerie Wood, Jeffrey L De Pons, Melinda R Dwinell, G Thomas Hayman, Mary L Kaldunski, Anne E Kwitek, Stanley J F Laulederkind, Marek A Tutaj, Mahima Vedi, Shur-Jen Wang, Peter D'Eustachio, Lucila Aimo, Kristian Axelsen, Alan Bridge, Nevila Hyka-Nouspikel, Anne Morgat, Suzi A Aleksander, J Michael Cherry, Stacia R Engel, Kalpana Karra, Stuart R Miyasato, Robert S Nash, Marek S Skrzypek, Shuai Weng, Edith D Wong, Erika Bakker, Tanya Z Berardini, Leonore Reiser, Andrea Auchincloss, Kristian Axelsen, Ghislaine Argoud-Puy, Marie-Claude Blatter, Emmanuel Boutet, Lionel Breuza, Alan Bridge, Cristina Casals-Casas, Elisabeth Coudert, Anne Estreicher, Maria Livia Famiglietti, Marc Feuermann, Arnaud Gos, Nadine Gruaz-Gumowski, Chantal Hulo, Nevila Hyka-Nouspikel, Florence Jungo, Philippe Le Mercier, Damien Lieberherr, Patrick Masson, Anne Morgat, Ivo Pedruzzi, Lucille Pourcel, Sylvain Poux, Catherine Rivoire, Shyamala Sundaram, Alex Bateman, Emily Bowler-Barnett, Hema Bye-A-Jee, Paul Denny, Alexandr Ignatchenko, Rizwan Ishtiaq, Antonia Lock, Yvonne Lussi, Michele Magrane, Maria J Martin, Sandra Orchard, Pedro Raposo, Elena Speretta, Nidhi Tyagi, Kate Warner, Rossana Zaru, Alexander D Diehl, Raymond Lee, Juancarlos Chan, Stavros Diamantakis, Daniela Raciti, Magdalena Zarowiecki, Malcolm Fisher, Christina James-Zorn, Virgilio Ponferrada, Aaron Zorn, Sridhar Ramachandran, Leyla Ruzicka, and Monte Westerfield. The Gene Ontology knowledgebase in 2023. *Genetics*, 224(1):iyad031, May 2023. ISSN 1943-2631. doi: 10.1093/genetics/iyad031. URL https://doi.org/10.1093/genetics/iyad031.

Sebastian Köhler, Michael Gargano, Nicolas Matentzoglu, Leigh C Carmody, David Lewis-Smith, Nicole A Vasilevsky, Daniel Danis, Ganna Balagura, Gareth Baynam, Amy M Brower, Tiffany J Callahan, Christopher G Chute, Johanna L Est, Peter D Galer, Shiva Ganesan, Matthias Griese, Matthias Haimel, Julia Pazmandi, Marc Hanauer, Nomi L Harris, Michael J Hartnett, Maximilian Hastreiter, Fabian Hauck, Yongqun He, Tim Jeske, Hugh Kearney, Gerhard Kindle, Christoph Klein, Katrin Knoflach, Roland Krause, David Lagorce, Julie A McMurry, Jillian A Miller, Monica C Munoz-Torres, Rebecca L Peters, Christina K Rapp, Ana M Rath, Shahmir A Rind, Avi Z Rosenberg, Michael M Segal, Markus G Seidel, Damian Smedley, Tomer Talmy, Yarlalu Thomas, Samuel A Wiafe, Julie Xian, Zafer Yüksel, Ingo Helbig, Christopher J Mungall, Melissa A Haendel, and Peter N Robinson. The Human Phenotype Ontology in 2021. *Nucleic Acids Research*, 49 (D1):D1207–D1217, January 2021. ISSN 0305-1048. doi: 10.1093/nar/gkaa1043. URL https://doi.org/10.1093/nar/gkaa1043.

Filip Ilievski, Daniel Garijo, Hans Chalupsky, Naren Teja Divvala, Yixiang Yao, Craig Rogers, Rongpeng Li, Jun Liu, Amandeep Singh, and Daniel Schwabe. KGTK: a toolkit for large knowledge graph manipulation and analysis. In *The Semantic Web–ISWC 2020: 19th International Semantic Web Conference, Athens, Greece, November 2–6, 2020, Proceedings, Part II 19*, pages 278–293. Springer, 2020.

L. Cappelletti, T. Fontana, E. Casiraghi, V. Ravanmehr, T.J. Callahan, C. Cano, M.P. Joachimiak, C.J. Mungall, P.N. Robinson, J. Reese, and G. Valentini. GRAPE for Fast and Scalable Graph Processing and random walk-based Embedding. *Nature Computational Science*, 2023. doi: 10.1038/s43588-023-00465-8.

Franz Baader, Ian Horrocks, and Ulrike Sattler. Description logics. *Foundations of Artificial Intelligence*, 3:135–179, 2008. ISBN: 1574-6526 Publisher: Elsevier.

Markus Krötzsch, Frantisek Simancik, and Ian Horrocks. Description Logics. *IEEE Intelligent Systems*, 29(1):12–19, 2014. doi: 10.1109/MIS.2013.123.

Yevgeny Kazakov, Markus Krötzsch, and František Simančík. The Incredible ELK: From Polynomial Procedures to Efficient Reasoning with Ontologies. *Journal of automated reasoning*, 53(1):1–61, 2014. ISBN: 0168-7433 Publisher: Springer.

David Osumi-Sutherland, Melanie Courtot, James P. Balhoff, and Christopher Mungall. Dead simple OWL design patterns. *Journal of Biomedical Semantics*, 8(1):18, June 2017. ISSN 20411480. doi: 10.1186/s13326-017-0126-0. URL https://jbiomedsem.biomedcentral.com/articles/10.1186/s13326-017-0126-0. Publisher: BioMed Central Ltd.

Nicolas Matentzoglu, Damien Goutte-Gattat, Shawn Zheng Kai Tan, James P. Balhoff, Seth Carbon, Anita R. Caron, William D. Duncan, Joe E. Flack, Melissa Haendel, Nomi L. Harris, William R. Hogan, Charles Tapley Hoyt, Rebecca C. Jackson, HyeongSik Kim, Huseyin Kir, Martin Larralde, Julie A. McMurry, James A. Overton, Bjoern Peters, Clare Pilgrim, Ray Stefancsik, Sofia MC Robb, Sabrina Toro, Nicole A. Vasilevsky, Ramona Walls, Christopher J. Mungall, and David Osumi-Sutherland. Ontology Development Kit: a toolkit for building, maintaining, and standardising biomedical ontologies. *Database*, 2022:baac087, October 2022. ISSN 1758-0463. doi: $10.1093/\mathrm{database}/\mathrm{baac087}$. URL http://arxiv.org/abs/2207.02056. arXiv:2207.02056 [cs].

Robert Arp and Barry Smith. Function, Role and Disposition in Basic Formal Ontology. In *Nature Precedings*. 2008.

Anita Bandrowski, Ryan Brinkman, Mathias Brochhausen, Matthew H. Brush, Bill Bug, Marcus C. Chibucos, Kevin Clancy, Mélanie Courtot, Dirk Derom, and Michel Dumontier. The ontology for biomedical investigations. *PloS one*, 11(4):e0154556, 2016. Publisher: Public Library of Science San Francisco, CA USA.

Clifford R. Jack Jr., David A. Bennett, Kaj Blennow, Maria C. Carrillo, Billy Dunn, Samantha Budd Haeberlein, David M. Holtzman, William Jagust, Frank Jessen, Jason Karlawish, Enchi Liu, Jose Luis Molinuevo, Thomas Montine, Creighton Phelps, Katherine P. Rankin, Christopher C. Rowe, Philip Scheltens, Eric Siemers, Heather M. Snyder, Reisa Sperling, Contributors, Cerise Elliott, Eliezer Masliah, Laurie Ryan, and Nina Silverberg. NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. *Alzheimer's & Dementia*, 14(4):535–562, 2018. ISSN 1552-5279. doi: $10.1016/\mathrm{j.jalz.2018.02.018}$. URL https://onlinelibrary.wiley.com/doi/abs/10.1016/j.jalz.2018.02.018. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1016/j.jalz.2018.02.018.

Yevgeny Kazakov. Modular Reuse of Ontologies: Theory and Practice. *JAIR*, 31, 2008. Publisher: JAIR.

Rebecca C. Jackson, James P. Balhoff, Eric Douglass, Nomi L. Harris, Christopher J. Mungall, and James A. Overton. ROBOT: a tool for automating ontology workflows. *BMC bioinformatics*, 20:1–10, 2019. Publisher: Springer.

Richard H. Scheuermann, Werner Ceusters, and Barry Smith. Toward an ontological treatment of disease and diagnosis. *Summit on translational bioinformatics*, 2009:116, 2009. Publisher: American Medical Informatics Association.

Santiago Timón Reina, M. Rincón Zamorano, and Atle Bjørnerud. Towards an Integrated Semantic Framework for Neurological Multidimensional Data Analysis. pages 175–184.

2015. doi: 10.1007/978-3-319-18914-7_18. URL http://link.springer.com/10.1007/978-3-319-18914-7_18.

B.-E. Kirsebom, R. Espenes, E. Hessen, K. Waterloo, S. Harald Johnsen, E. Gundersen, S. Botne Sando, G. Rolfseng Grøntvedt, S. Timón, and T. Fladby. Demographically adjusted CERAD wordlist test norms in a Norwegian sample from 40 to 80 years. *Clinical Neuropsychologist*, 2019. ISSN 17444144. doi: 10.1080/13854046.2019.1574902.

J. Espenes, E. Hessen, I.V. Eliassen, K. Waterloo, M. Eckerström, S.B. Sando, S. Timón, A. Wallin, T. Fladby, and B.-E. Kirsebom. Demographically adjusted trail making test norms in a Scandinavian sample from 41 to 84 years. *Clinical Neuropsychologist*, 2020. ISSN 17444144. doi: 10.1080/13854046.2020.1829068.

Ingrid Myrvoll Lorentzen, Jacob Espenes, Erik Hessen, Knut Waterloo, Geir Bråthen, Santiago Timón, Dag Aarsland, Tormod Fladby, and Bjørn Eivind Kirsebom. Regression-based norms for the FAS phonemic fluency test for ages 40–84 based on a Norwegian sample. *Applied Neuropsychology:Adult*, 2021. ISSN 23279109. doi: 10.1080/23279095.2021.1918128. URL https://www.tandfonline.com/doi/abs/10.1080/23279095.2021.1918128. Publisher: Routledge.

Ingvild Vøllo Eliassen, Bjørn-Eivind Kirsebom, Tormod Fladby, Sigrid Botne Sando, Mathilde Suhr Hemminghyth, Dag Aarsland, Santiago Timón-Reina, Anders Wallin, Fredrik Öhman, and Marie Eckerström. Regression-based cognitive change norms applied in biochemically defined predementia Alzheimer's disease. *Neuropsychology*, 2022. Publisher: American Psychological Association.

Jacob Espenes, Ingvild Vøllo Eliassen, Fredrik Öhman, Erik Hessen, Knut Waterloo, Marie Eckerström, Ingrid Myrvoll Lorentzen, Cecilie Bergland, Madelene Halvari Niska, and Santiago Timón-Reina. Regression-based normative data for the Rey Auditory Verbal Learning Test in Norwegian and Swedish adults aged 49–79 and comparison with published norms. *The Clinical Neuropsychologist*, pages 1–25, 2022. Publisher: Taylor & Francis.

David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, Nazanin Assempour, Ithayavani Iynkkaran, Yifeng Liu, Adam Maciejewski, Nicola Gale, Alex Wilson, Lucy Chin, Ryan Cummings, Diana Le, Allison Pon, Craig Knox, and Michael Wilson. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research*, 46(D1): D1074–D1082, January 2018. ISSN 0305-1048. doi: 10.1093/nar/gkx1037. URL http://academic.oup.com/nar/article/46/D1/D1074/4602867. Publisher: Oxford University Press.