UNIVERSIDAD POLITÉCNICA DE MADRID
Escuela Técnica Superior de Ingeniería Agronómica, Alimentaria y de Biosistemas

Applying deep semantics to the representation of clinical data to improve machine usability

DOCTORAL THESIS

Submitted for the degree of Doctor by:

Pablo Alarcón Moreno
BSc in Biochemistry and Molecular Biology
MSc in Bioinformatics

Madrid, 2024

UNIVERSIDAD POLITÉCNICA DE MADRID
ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA AGRONÓMICA, ALIMENTARIA Y DE BIOSISTEMAS

Doctoral Degree in Biotechnology and Genetic Resources of Plants and Associated Microorganisms

Applying deep semantics to the representation of clinical data to improve machine usability

DOCTORAL THESIS

Submitted for the degree of Doctor by:

Pablo Alarcón Moreno
BSc in Biochemistry and Molecular Biology
MSc in Bioinformatics

Under the supervision of:
Dr. Mark Denis Wilkinson

Madrid, 2024

Title: Applying deep semantics to the representation of clinical data to improve machine usability

Author: Pablo Alarcón Moreno

Doctoral Programme: Biotechnology and Genetic Resources of Plants and Associated Microorganisms

Thesis Supervision: Dr. Mark Denis Wilkinson, Senior Researcher, Escuela Técnica Superior de Ingeniería Agronómica, Alimentaria y de Biosistemas, Universidad Politécnica de Madrid. (Supervisor)

External Reviewers:

Thesis Defense Committee:

Thesis Defense Date:

This work was supported by the European Joint Programme on Rare Diseases (EJP RD) and the International Rare Diseases Research Consortium (IRDiRC). The European Joint Programme on Rare Diseases, including the IRDiRC Scientific Secretariat is funded by the European Union under the European Union's Horizon 2020 research and innovation programme Grant Agreement N°825575.

Acknowledgement

“There is a curse. They say:

May you live in interesting times.” — Terry Pratchett

Grandes momentos y muchos motivos me han llevado aquí, y muchos de esos motivos tienen nombre y apellido.

A toda mi familia, que os quiero mucho. Aunque no esté allí con vosotros, porque la vida siempre me tira lejos de los huertos en los que me crie. No olvidéis que allá donde voy, venís conmigo a todos estos lugares. A mis abuelos que el orgullo en sus miradas me ha hecho seguir adelante en los momentos más duros, y porque nunca olvido de dónde vengo. Espero que nunca dejéis de sentiros orgullosos de mí. Abuelo, sé que nunca dejaste de estar hasta el final. Esta tesis que por poco no pudiste ver terminada, te la dedico a ti. A mis padres que, sin su sacrificio, desde el primer día hasta el último, siempre me apoyasteis en todas las decisiones que tomé. Vosotros me disteis la vida y la oportunidad de estudiar. Nunca olvidaré este sacrificio, y espero que también os sintáis orgullosos de todo mi camino recorrido. A mis hermanos que, aunque la distancia nos mantenga alejados, cada vez esta será más pequeña. No os olvidéis de vuestro hermano mayor y de todos los momentos compartidos, siempre os llevo en mi corazón, esta tesis os la dedico a vosotros.

To my wonderful working team at Centro de Biotecnología y Genómica de Plantas. Alberto, Oussama, thank you so much for your unconditional support in the darkest moment of high stress, and for the joy in the brightest celebrations. Looking forward to celebrating this thesis with you. To my supervisor Mark, for its wonderful support and for the opportunity of working in this amazing field that great minds like yours have created. Hope we ring the bell as hard as we can once this thesis is finished. I can never thank you enough. A Marina que, sin tu apoyo y ayuda incondicional durante todo el programa de doctorado. Sin la menor de las dudas, no seguiría aquí si no fuese por ti. Gracias por ayudarme tanto y por haber estado ahí cuando el pánico afloraba.

A Mireia, que estuviste al comienzo de este camino y me enseñaste a esforzarme, en cada minuto y confiar en mí mismo, me pregunto qué habría sido de mi sin esa entereza que me has enseñado. Esta tesis lleva tu nombre tatuado. A mi familia

elegida, a mis amigos en Murcia y Madrid, que me han acogido y me han escuchado cuando los momentos han sido duros, os agradezco tanto a todos que esta tesis no habría salido sin vuestro apoyo. A todos, estéis en estas páginas o no, no habéis hecho más que cuidarme hasta llegar a este momento. Por todos vosotros estoy aquí y esta tesis es vuestra. A Adri, por ser mi hermano allí a donde voy, llevas a mi lado tantos años que ya no los cuento, y todos ellos siempre has estado en lo fácil y lo difícil. Tú que me has visto cambiar cien veces, que me has visto caer y levantarme, que me has visto hacerme pequeño y grande con esta tesis y con tantas otras cosas. Lo conseguí, y espero que te sientas orgulloso porque sin ti no habría llegado hasta aquí. A Bruno, que en tan poco tiempo te has vuelto imprescindible, que tu apoyo cuando estoy perdido me resulta fundamental. Espero que celebremos el final de esta tesis como si fuese nuestra, y sé que lo haremos. A Alex por enseñarme la importancia de los cuidados a los que queremos. A Quique por hacerme soñar en lugares mágicos cuando la crudeza del día a día se hace cuesta arriba. A Marta, por todos los momentos compartidos, tú que me has escuchado y me has apoyado en estos momentos tan complicados haciéndome recordar la importancia de la salud mental en los días más raros.

A Dani y a Mariaje, por ser papá y mamá en la calle Ave María, por no esperar nada cambio y siempre cuidarme, sé que siempre os tengo cuando me necesitáis. Hemos evolucionado tanto desde que nos conocimos, pero todo este tiempo la vida nos ha llevado hasta aquí. Estoy deseando ver las nuevas cosas que nos traerán juntos. A Claudia, Carmen, Lota, Chuck, Fayna, Mery, Andrea y Alicia, que me habéis abierto un hueco en vuestra Madrid. Ahora es nuestra Madrid, aunque las marquesas hayáis colonizado nuevos lugares, pero ahora todos esos lugares son casa. Gracias por hacerme un hueco, gracias por haberme hecho sentir tan vivo, no olvidaré ese cariño tan incondicional, esta tesis es nuestra. A Celia, que hemos llegado hasta aquí, en una ciudad que nunca fue nuestra pero ahora está llena de chinchetas para el recuerdo. Gracias por leerme la mente a cada momento sabiendo cuando necesito un abrazo infinito. Por hacerme sobrevivir a los berenjenales de esta tesis y, sobre todo, todo. A Mar, que me has enseñado la bondad y la sencillez de ser uno mismo sin máscaras. Gracias por recordarme que lo maravilloso de las cosas buenas, que el bien gana, y nosotros con él, si queremos.

María, ahora te toca a ti.

Abstract

In the healthcare domain, particularly within the rare disease community, there is a notable increase in the volume and variety of clinical information. Recent activities and efforts coming from the European Joint Programme on Rare Diseases (EJP-RD) aim to address this challenge by the use of Linked Data and Semantic Web technologies that enable descriptions of data in a machine-interpretable manner. Adhering to the FAIR data principles (Findable, Accessible, Interoperable, and Reusable), the objective of this thesis is to use this large-scale ongoing initiative to test the validity of the claim that FAIR data and semantic technologies lead to increased interoperability and machine-actionability.

In an attempt to harmonize the data landscape among the EJP-RD partners, a process of data modelling was undertaken, starting from a defined set of common data elements enumerated by the European Platform on Rare Disease (EU RD) Platform, and expanding that model with more data elements identified in European Rare Disease patient registries. This effort led to the creation of the Common Data Elements Semantic Model (CDE-SM) along with a set of Semantic Web services for data pre-evaluation and transformation, utilizing YARRRML and CSV templates.

Preliminary studies and experiments concerning data interoperability were done by integrating the CDE-SM with other standardized data models present in the healthcare community, such as Biolink Model and C-PATH. By leveraging the Biolink model to bridge between CDE-SM and the C-PATH data models, common SPARQL queries were formulated to identify and query shared structures across both models, increasing schema harmonization at the data querying level to a limited extent.

The semantic data model was extended to cover a wider range of data types, and was renamed to the Clinical and Registry Entries Semantic Model (CARE-SM). Consistency between data elements' representations allowed several implementation improvements, including simplified data transformation, improved data discoverability, and deployment of a Beacon API service, enabling anonymous, federated querying and aggregation of patient data. Additionally, these improvements facilitated the conversion of data represented using the CARE-SM model to data compliant with the OMOP-CDM through the creation of

a schema mapping between these data models and the implementation of an Extract, Transform, Load (ETL) workflow.

This thesis demonstrates the successful interoperability by implementing CARE-SM through the deployment of the Beacon API. This success is attributed to the use of a common structure and shared vocabularies, facilitating interoperability. However, our experiments revealed that the creation of a FAIR data model did not significantly enhance interoperability with other standards. The primary advantages of CDE-SM and CARE-SM lie in their structural consistency, showing how FAIR data is necessary but not sufficient for achieving interoperability.

Resumen

En el ámbito de la salud, particularmente dentro de la comunidad de enfermedades raras, hay un notable aumento en el volumen y la variedad de información clínica. Las actividades y esfuerzos recientes provenientes del European Joint Programme on Rare Diseases (EJP-RD) tienen como objetivo abordar este desafío mediante el uso de Datos Enlazados y tecnologías de la Web Semántica que permiten la descripción de datos de manera interpretable por máquinas. Adhiriéndose a los principios de datos FAIR (Findable, Accessible, Interoperable, and Reusable), el objetivo de esta tesis es utilizar esta iniciativa en curso a gran escala para probar la validez de la afirmación de que los datos FAIR y las tecnologías semánticas conducen a una mayor interoperabilidad y capacidad de acción por parte de las máquinas.

En un intento por armonizar el panorama de datos entre los socios del EJP-RD, se llevó a cabo un proceso de modelado de datos, comenzando con un conjunto de elementos de datos comunes enumerados por la Plataforma Europea sobre Enfermedades Raras (EU RD Platform) y ampliando ese modelo con más elementos de datos, identificados en los registros de pacientes con enfermedades raras de Europa. Este esfuerzo llevó a la creación del Modelo Semántico de Elementos de Datos Comunes (CDE-SM, por sus siglas en inglés), junto con un conjunto de servicios de la Web Semántica para la pre-evaluación y transformación de datos, utilizando plantillas YARRRML y CSV.

Se realizaron estudios y experimentos preliminares sobre la interoperabilidad de datos integrando el CDE-SM con otros modelos de datos estandarizados presentes en la comunidad de salud, como el Modelo Biolink y C-PATH. Al aprovechar el modelo Biolink para crear un puente entre CDE-SM y el modelo de datos C-PATH, se formularon consultas SPARQL comunes para identificar y consultar estructuras compartidas entre ambos modelos, aumentando la armonización del esquema a nivel de consulta de datos en cierta medida.

El modelo de datos semánticos se extendió para cubrir una gama más amplia de tipos de datos y se renombró como Modelo Semántico de Entradas de Registros Clínicos (CARE-SM, por sus siglas en inglés). La consistencia entre las representaciones de los elementos de datos permitió varias mejoras en la implementación, incluyendo la simplificación de la transformación de datos, la mejora del descubrimiento de datos y el despliegue de un servicio API Beacon, que permite la consulta federada anónima y la agregación de datos de pacientes.

Además, estas mejoras facilitaron la conversión de datos representados utilizando el modelo CARE-SM a datos compatibles con OMOP-CDM mediante la creación de un mapeo de esquemas entre estos modelos de datos y la implementación de un flujo de trabajo de Extracción, Transformación y Carga (ETL).
Esta tesis demuestra la interoperabilidad exitosa mediante la implementación de CARE-SM a través del despliegue de la API Beacon. Este éxito se atribuye al uso de una estructura común y vocabularios compartidos, facilitando la interoperabilidad. Sin embargo, nuestros experimentos revelaron que la creación de un modelo de datos FAIR no mejoró significativamente la interoperabilidad con otros estándares. Las principales ventajas de CDE-SM y CARE-SM radican en su consistencia estructural, demostrando cómo los datos FAIR son necesarios, pero no suficientes para lograr la interoperabilidad.

Table of Contents

List of Figures

List of Tables

Abbreviations and Acronyms

API	Application Programming Interface
BFO	Basic Formal Ontology
C-Path	Critical Path Institute
CDE	Common Data Element
DQD	Data Quality Dashboard
EC RD Platform	European Platform on Rare Disease
EJP-RD	European Joint Programme on Rare Diseases
EMA	European Medicines Agency
ERKreg	European Rare Kidney Disease Reference Network
ERN	European Reference Network
EU	European Union
FAIR	Findable, Accessible, Interoperable and Reusable
FDA	Food and Drug Administration
GUID	Global Unique Identifier
HGNC	HUGO Gene Nomenclature Committee
HGVS	Human Genome Variation Society
HPO	Human Phenotype Ontology
IRI	Internationalized Resource Identifier
JSON	JavaScript Object Notation
NCIT	National Cancer Institute Thesaurus
OBO	Open Biological and Biomedical Ontology
OHDSI	Observational Health Data Sciences and Informatics
OMIM	Online Mendelian Inheritance in Man
OMOP-CDM	Observational Medical Outcomes Partnership Common Data Model
ORDO	Orphanet Rare Disease Ontology

PKD	Polycystic Kidney Disease
PRO	Patient Reported Outcome
PROV-O	PROVenance Ontology
R2RML	RDB to RDF Mapping Language
RDF	Resource Description Framework
RDFS	Resource Description Framework Schema
REST	Representational State Transfer
RML	RDF Mapping Language
SDTM	Study Data Tabulation Model
SHACL	Shapes Constraint Language
ShEx	Shape Expressions
SIO	Semanticscience Integrated Ontology
SKOS	Simple Knowledge Organization System
SNOMED-CT	Systematized Nomenclature of Medicine – Clinical Terms
SPARQL	SPARQL Protocol and RDF Query Language
SQL	Structured Query Language
SSSOM	Simple Standard for Sharing Ontology Mappings
UPM	Universidad Politécnica de Madrid
URI	Uniform Resource Identifiers
W3C	World Wide Web Consortium
XML	Extensible Markup Language
YAML	Yet Another Markup Language

1. Introduction

In the era of big data, the volume and variety of data generated in the healthcare domain has increased rapidly. Historically captured in legacy formats such as paper-based documents, data practices in healthcare have evolved to become digitized; however, this has not been accompanied by an enhancement of reusability. Healthcare data is most often structured in table-based formats that don't contain enough context for machine agents to interpret their meaning, thus data sharing and integration can only be achieved manually. Traditional approaches to data sharing often struggle with the scale and complexity of modern data ecosystems, leading to inefficiencies, siloed datasets, and barriers to data reuse.

The advent of machine-readable information in the past 20 years holds the promise of better data management, traffic and interpretation. In 2016, (Wilkinson et al., 2016), laid-out a set of patterns for data and metadata publication aimed at creating "an internet of FAIR data and services". These guidelines were named the FAIR principles. These principles primarily focus on and emphasize the importance of data discovery, access, and reuse by machines.

One of the most challenging of the FAIR principles are these related to "I" - Interoperability - and particularly with respect to the reuse of data by machines. Achieving interoperability requires more than just technical standards or infrastructure; it demands that the computational agent "understand" the data it has discovered to a sufficient degree that it can correctly reuse it. It requires data to be not only machine readable, but machine interpretable. Semantic Web technologies and Linked data practices play a critical role, providing frameworks and standards for both human and machine data interpretability, and are at the core of the FAIR initiative.

There has been a notable increase in the number of highly specialized patient registries during the last decades as diagnostic tools have become capable of differentiating many distinct but similar disorders. This has worsened the integration problem by fragmenting and distributing the data even more, with each site usually having its own non-machine-readable data model. This lack of

data interoperability is causing researchers to invest valuable time - up to 80% of their total data-focused investment - finding, preparing, filtering, and combining datasets (European Commission. Directorate General for Research and Innovation. & PwC EU Services., 2018).

One of the domains of knowledge that have suffered from these difficulties in particular is the Rare disease (RD) community. Rare diseases affect a small percentage of the population compared with high-incidence diseases like diabetes or cancer. Although a single rare disease may not be prevalent globally, the collective impact of all rare diseases affects millions of people. Up to 36 million people are affected by a rare disease in the EU, which is approximately 8% of the total EU population (Damme et al., 2023). Patient rare disease data is geographically fragmented and scarce. Their information is spread over many "boutique" repositories, often curated under site-specific protocols or standards and often lacking sufficiently descriptive controlled vocabularies or ontologies due to the rarity of the disorder and its phenotypic effects.

The EU, in its attempt to deal with this fragmented and scarce data distribution, has set up European Reference Networks (ERN). These ERNs gather several of these data silos dedicated to a particular domain of healthcare data into a network for exchanging knowledge and information among health care providers. These networks aim to improve access to common and precise clinical observations for the rare disease community in Europe.

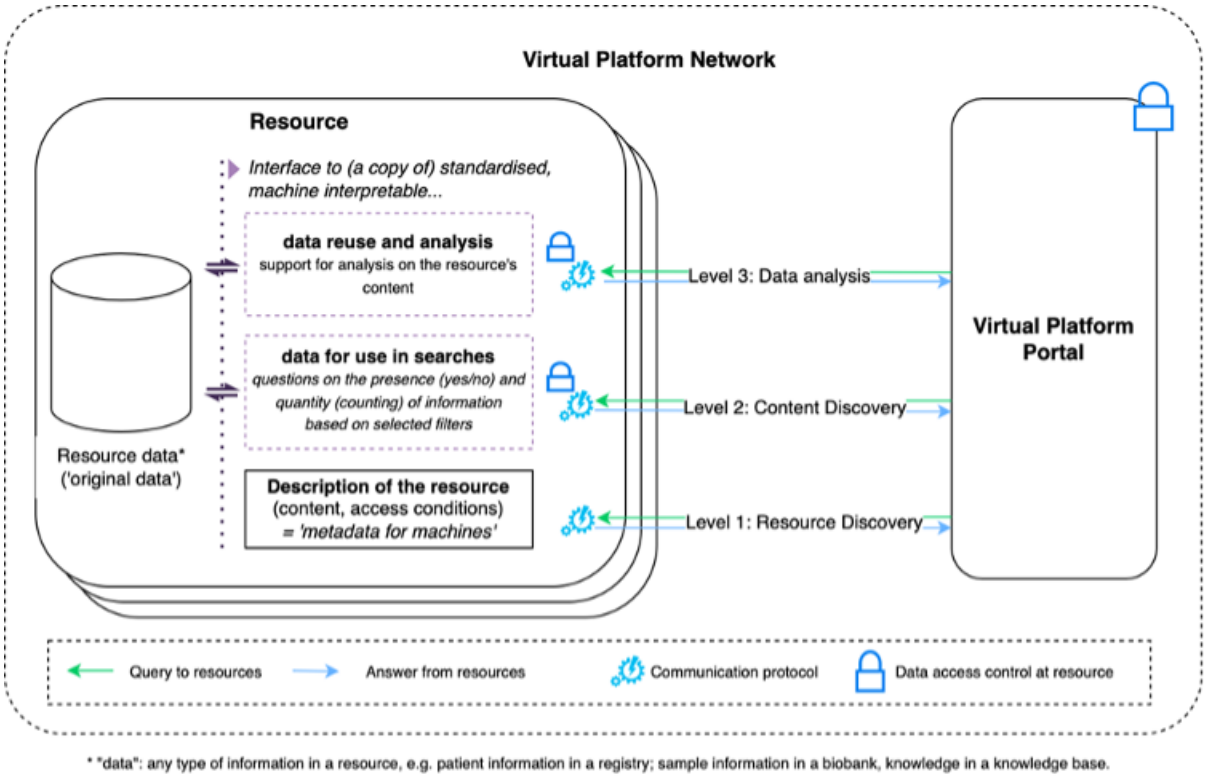


Figure 1: Overview of the planned functionality of the EJP-RD Virtual Platform.

Along with the established ERNs, the Horizon 2020 initiative funded the creation of The European Joint Programme on Rare Diseases (EJP-RD), a project aimed to address the challenges faced by the rare disease community, including:

- Integrating diverse stakeholders from various countries, including researchers, healthcare providers, and patient organizations as collaborators, to work towards encouraging innovation in rare disease research

- Ensuring data sharing and interoperability from different institutions and countries. Making sure this data is compatible and can be shared effectively, maintaining privacy and security.

- Securing sufficient and sustainable funding to support research, clinical trials, and other activities essential for progress in rare disease research.

In order to achieve data sharing and interoperability, EJP-RD has as an objective the creation and deployment of a Virtual platform (Figure 1) capable of performing federated discovery and analysis of heterogeneous rare disease data. The Virtual

Platform would leverage Semantic Web technologies to achieve this goal. FAIR data principles form the basis of this objective, defining a guideline for machines to interpret the conceptual representation of this complex domain of knowledge.

EC RD Platform defined a set of minimal data entries present in most of the patient data registries throughout Europe, with this list being referred-to as the Common Data Elements (CDE). This list includes the permitted vocabularies for its annotation, data types and the relationship between some of this element, grouped under a common logic. These CDEs became the first target for creating a semantic data model capable of representing patient data registries.

As it's addressed above, one of the main goals of the EJP-RD project is to ensure data sharing and interoperability. The different standards present at all these institutions is a challenge for achieving our goal in data integration. Some of these institutions have already adopted healthcare standards for data modelling and management in a semantic manner. Data interoperability activities are being coordinated, offering a significant opportunity to explore new approaches from different perspectives and methodologies. This Thesis explores several of these methodologies.

2. Objectives and research questions

The main objective of this Thesis is to explore the degree to which adding "deep semantics" to clinical research data, and closely following the FAIR principles, would lead to interoperability between clinical information systems. We will ask (and answer) the question of whether adherence to the Principles is sufficient to provide the machine-reusability and interoperability that were promised in the seminal FAIR publication. This thesis had the following specific research questions:

- Is it possible to create a generalized semantic data model within the domain of rare diseases, using the Common Data Elements for Rare disease registration as an initial target?
- Is this generalized model scalable for future expansion beyond the CDEs?
- Can we perform federated data exploration using our semantic data model?
- Does the semantic model facilitate interoperability with other common clinical data frameworks?

3. Background

3.1. Bioinformatics (Biological Informatics)

The precise definition of bioinformatics is a matter of some debate, but the history of this evolving field begins in the 1970s, defined as "the study of informatic processes in biotic systems" (Hogeweg, 2011). Through time, bioinformatics has evolved to become a heterogenous field that covers multiple disciplines, from computational modelling to structural biology and genomics (Bayat, 2002; Luscombe et al., 2018) . Exponential expansion in the quantity of molecular data, especially with the rise of high-throughput technologies dedicated to DNA and protein sequencing (Altman et al., 1999; BergerBonnie et al., 2016), has driven bioinformatics into the "omics" era, with sub-fields such as genomics, proteomics or transcriptomics. By the use

of computational tools and frameworks, bioinformatics organizes, analyzes and interprets big data, offering insights into complex biological pathways, phenotypes, genotypes and disease mechanisms (Gómez-López et al., 2019). The incorporation of bioinformatics into the healthcare domain has resulted in new domains of exploration such as personalized medicine (Sunil Krishnan et al., 2021).

3.2. The World Wide Web (WWW)

The World Wide Web, also called the Web or W3, is a set of standards for information publishing allowing accessibility and sharing of content over the Internet. These standards were invented by Sir Tim Berners-Lee in 1989 and since then have revolutionized the way both humans and machines communicate and consume information (Raffl et al., 2008).

The Web relies on several core components:

- The Hypertext Transfer Protocol (HTTP) defines the structure of the request (client-side) and response (server-side) messages that facilitate the sharing of HTML (and other) documents.
- URL (Uniform Resource Locator) or URI (Uniform Resource Identifier) are global unique identifiers that a client uses to access Web components. Both URL and URI are formally referred to as Universal Document Identifiers. The thing identified by a URL/URI is commonly referred to as a "Resource", and this term will be used throughout this thesis.
- The Hypertext Markup Language (HTML) is the standard syntax used for publishing documents on the Web. It structures Web content and facilitates the creation of Web pages. HTML enables linking documents one to another by using "hyperlinks". These hyperlinks facilitate navigation across the internet.

The World Wide Web Consortium (W3C) was also founded by Tim Berners-Lee in 1994 (Jacobs, 2023.). The W3C is in charge of defining and standardizing new technologies intended to be used by Web applications.

3.3. The Semantic Web (Web 3.0)

The Semantic Web is an extension of the World Wide Web, capable not only of interlinking data within the Web, but also assisting machines in understanding and interpreting this information. The core technology used to build the Semantic Web - Linked Data - was proposed by Tim Berners-Lee in the late 1990s and (informally) published in 2001

Resource Description Framework (RDF) - a W3C recommended standard since 2004 - is a flexible framework used for data and metadata exchange, and is what allows the creation of Linked data.

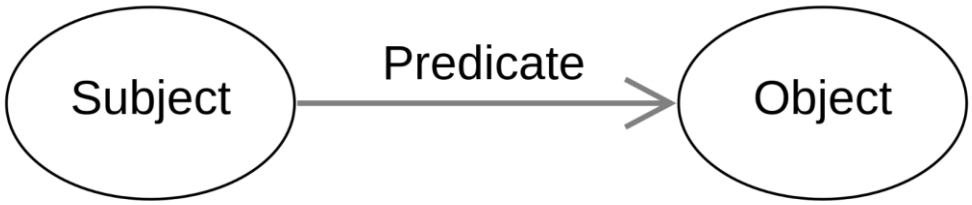


Figure 2: Visual representation of an RDF-Triple.

RDF provides a framework for describing and interlinking structured data, facilitating the sharing and reuse of this information. RDF structures the information as a collection of "triples" (Figure 2), each consisting of three components referred-to as the "subject", "predicate", and" object", where the predicate indicates the nature of the relationship between the subject and the object. In more recent iterations of the RDF specification[1] a fourth element can be included called the "context" or "named graph" and provides a way to group RDF triples. The resulting structure, called an RDF-Quad, allows for more complex data modeling and the possibility to specify additional information such as provenance or the scope of a set of triples.

RDF is capable of representing a wide range of data and metadata information, including Web resources and structured databases. RDF uses URIs to uniquely identify the Web resources and the relationships that are part of the RDF structure, enabling interlinking resources within the Web, even if these resources are not located in the same Web document. These capabilities facilitate data

[1] https://www.w3.org/TR/rdf12-n-quads/

integration at the protocol level, but do not in themselves contribute to interoperability - i.e., a computational agent engaged in integrating data via these linkages does not have a way to know that it is doing so correctly or in any meaningful way.

A second W3C standard that addresses this problem is the Web Ontology Language (OWL) (McGuinness & Van Harmelen, 2004). OWL is a Description Logic (DL) that evolved as a combination of two previous Description Logics - DAML and OIL (Horrocks, 2002). This DL is used in the formal descriptions and restrictions of concepts in a particular domain, and the relationship(s) between them. OWL includes the concept of a sub-class, allowing concepts and relationships to be hierarchically defined, with child concepts inheriting the logical descriptions of their parent concepts.

OWL is essential for the definition of ontologies that contain logical and hierarchical definitions. Ontologies provide a shared and common vocabulary for describing concepts and their relationships in a specific domain. Ontologies empower data integration and interoperability by assisting machines' interpretation of data represented in Linked Data.

For exploration of Linked Data, the SPARQL Protocol and RDF Query Language (SPARQL) allows queries to be represented and exchanged. This language provides a standardized syntax for creating queries over RDF data in the form of graph patterns. SPARQL servers that host RDF data (called "triplestores") will identify any graph in the triplestore that matches the incoming query pattern and return any bound variables in the query request. SPARQL also defines the request and response messages that can be passed between client and server (Hommeaux & Seaborne, 2008; Pérez et al., 2009).

Despite the Semantic Web's numerous advantages, adoption has been slow, lacking a critical mass of Web content that has been semantically annotated and difficulty of adding metadata and structured data to existing Web content (Beno et al., 2019; Manuja & Garg, 2011). However, as more and more organizations adopt semantic technologies, and as more data is made available in structured form, the Semantic Web is likely to become more widely adopted in the years to come.

3.4. 5-star Linked Data

The 5 Star Linked Data model[2] , proposed by Tim Berners-Lee, suggested a 5-star evaluation system for Linked Data. This system is accumulative, and each additional star expected the previous star is already achieved. This 5-stars criteria enumerated here:

1. Data is available on the Web, in whatever format.

2. Available as machine-readable structured data, (i.e., not a scanned image).

3. Available in a non-proprietary format, (i.e, CSV, not Microsoft Excel).

4. Published using open standards from the W3C (RDF and SPARQL).

5. All of the above and links to other Linked Open Data.

Linked data is a set of practices for connecting structured data on the Web (Bizer et al., 2009). The use of Linked data is defined by four guidelines that describes the Linked data paradigm[3]:

- Use URIs to identify resources. By using a unique URI for assigning every Web document, ensure that every resource in the Web of data is findable by machines.

- Use HTTP URIs so that people can look up those resources via standard Web protocols

- When someone looks up a URI, provide useful information, using standards like RDF and SPARQL. This facilitates the sharing of common vocabulary and its relationship.

[2] https://www.w3.org/2011/gld/wiki/5_Star_Linked_Data

[3] https://www.w3.org/DesignIssues/LinkedData.html

- Include links to other URIs so that they can discover more resources, creating a network of interlinked information.

These key rules highlight the importance of using Semantic Web technologies to enhance machine data interpretability and accessibility. By adhering to this philosophy, the Web of data can become a rich, interconnected network space for data sharing.

3.5. SemanticScience Integrated Ontology (SIO)

The SemanticScience Integrated Ontology (SIO) is an ontology for representing and annotating data that emerges from biomedical research (Dumontier et al., 2014).

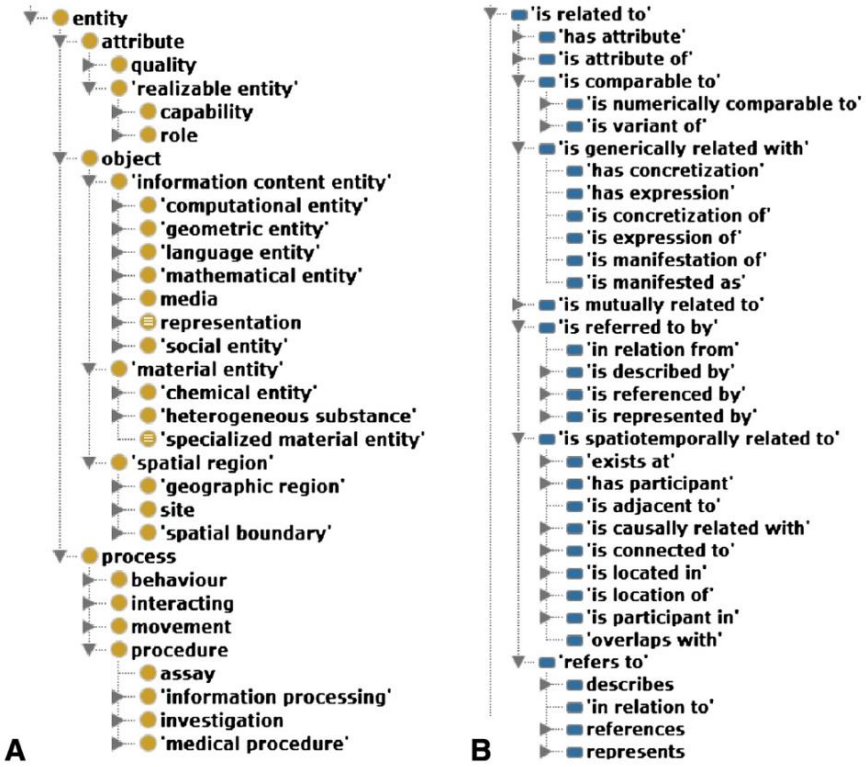


Figure 3: Selected portions of (A) class and (B) object property hierarchies in SIO.

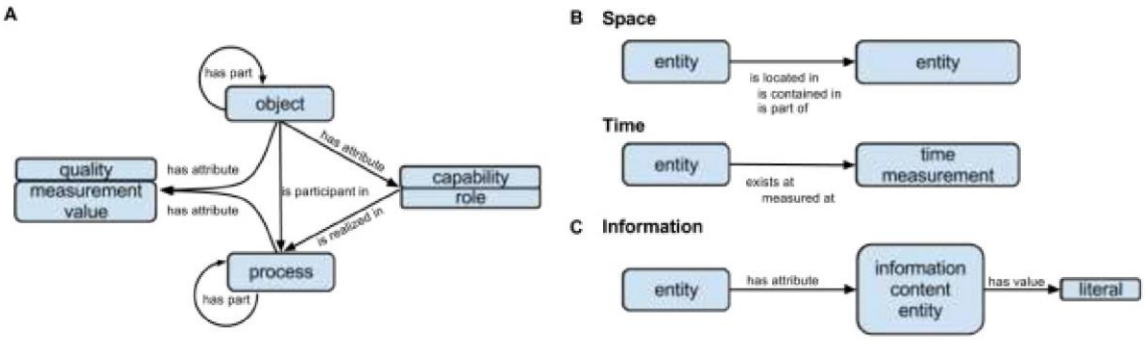


Figure 4: Key sub-components and relations in SIO.

Figure 4 shows a summary of the class and object-property hierarchies where 'entity' is the top-level class and 'is related to' is the top-level object property. Classes in SIO, inherited from entity, are structured as objects, processes and their attributes (qualities, capabilities, roles, measurement values):

- Objects, an entity that occupies space and is fully identifiable by its characteristics at any moment in time in which it exists. "Information content entity" is a child concept of Object. The value of this information entity is represented as literal, using the "has value" data property. (Figure 4.C)
- Process, being an entity that "unfolds in time and has temporal parts". Any attribute, such as the "role" exists at some time in the process that bears it, but it "is realized in" a process (Figure 4.A)
- Attribute, that gathers the concepts of 'quality' (intrinsic attribute), 'capability' (action specification) or 'role' (behavior, right and obligation).

Also, SIO offers a number of mereotopological relations that can be used to describe one or more entities in terms of their spatial organization, shown in Figure 4.B

3.6. OBO Foundry

The OBO Foundry is a collaborative effort aimed at developing a set of interoperable and reusable ontologies in the biomedical domain. The principles of the OBO Foundry emphasize openness, collaboration, and adherence to community standards. The National Cancer Institute Thesaurus (NCIT) is one of the ontologies included in the OBO catalog. NCIT provides a comprehensive vocabulary for describing both cancer and non-cancer-related concepts, including diseases, treatments, and biological processes, and is extensively used in this

thesis work. It serves as a valuable resource for biomedical research, facilitating data integration and interoperability in the field of oncology and general healthcare domain.

3.7. The FAIR Principles.

In January 2014 a group of hand-selected interested parties were invited to join Dr. Barend Mons - a biomedical informatics researcher in Leiden, NL - at a workshop aimed at tackling one of the "Grand Challenges of E-Science": the appropriate discovery and reuse of data by mechanized agents. From the workshop emerged the "FAIR Guiding Principles" - a set of practices and design patterns for publishing data that is Findable, Accessible, Interoperable and Reusable by both machines and humans (Wilkinson et al., 2017).

The FAIR Principles can be summarized as follows (quoted from Wilkinson et al., 2017) :

> ***Findable***—*data should be identified using globally unique, resolvable, and persistent identifiers, and should include machine-actionable contextual information that can be indexed to support human and machine discovery of that data.*
>
> ***Accessible***—*identified data should be accessible, optimally by both humans and machines, using a clearly-defined protocol and, if necessary, with clearly-defined rules for authorization/authentication.*
>
> ***Interoperable***—*data becomes interoperable when it is machine-actionable, using shared vocabularies and/or ontologies, inside of a syntactically and semantically machine-accessible format.*
>
> ***Reusable***—*Reusable data will first be compliant with the F, A, and I principles, but further, will be sufficiently well-described with, for example, contextual information, so it can be accurately linked or integrated, like-with-like, with other data sources. Moreover, there should be sufficiently rich provenance information so reused data can be properly cited.*

The concepts behind the FAIR Principles are not, in fact, novel. They are a consolidation and summarization of decades-old and well-established best practices in information science, library science, and computer science. The subsequent popularity of the FAIR Principles - now with nearly 14,000 citations and becoming core recommendations from journals and funding agencies worldwide - is at least in part due to the "friendly" acronym, and also in part due to the approachable summarization of these otherwise complex, deeply-studied, and interdisciplinary data sharing practices. Conversely, this simplification of complex ideas has led to broad re-/mis-interpretation of the intentions of the authors (Mons et al., 2017) and thus a wide range of organizations have claimed to "already be FAIR" as a result of flexible interpretations of what the Principles require of data publishers.

In this thesis we follow a strict interpretation of the FAIR Principles, as envisioned by the lead author. All technologies and solutions discussed in this work are compliant with the requirements of the Principles. As such, this thesis represents a challenge to the FAIR Principles.

3.8. Ontology mapping

Ontology mapping, also called as ontology matching or ontological alignment, establishes connections between similar concepts described in different ontologies, enabling data integration and interoperability between datasets that are annotated using different vocabularies. Ontology mapping is central to providing semantic access across aggregated data used in knowledge-based products and services consumed by life science companies, academic institutions, and universities (Harrow, 2019). There several techniques for creating mappings between ontologies:

- **Manual mapping:** Experts in the field manually create a mapping between at least two different concept representations, ensuring accuracy by analyzing the similarity between each concept representation in common between the ontologies.

- **Automatic or Semi-automatic mapping:** The use of algorithms to identify the potentially shared concepts between different ontologies This methodology

often utilizes machine learning and statistical methods to identify potential matches (Li et al., 2019; Zhou et al., 2020). An example of this mapping would be semantic similarity methodologies for quantifying how alike two concepts are based on their meaning. This method combines structural aspects of the graph with the lexical information content of the concepts (Zhu, 2017).

The Simple Knowledge Organization System[4] (SKOS), developed by W3C, is a standard vocabulary for mapping ontologies. The SKOS terminology includes concepts such a "broader than", "narrower than", and "exact match" that can be used to describe the similarity between two concepts across ontologies (Bueno-de-la-Fuente, 2008).

Ontology mapping faces several challenges. The heterogeneity of concept terminologies and the different granularities in concept definitions between ontologies make mapping both challenging, and often of limited utility due to low accuracy. Improving the reliability of mapping generally requires a post-validation technique, such as community feedback and crowdsourcing (Harrow, 2019; Li et al., 2019).

3.9. The Biolink Model

The Biolink Model is an open-source schema and data model designed by Chris Mungall et al. Biolink standardizes entity-relationships from the biomedical and molecular biology domains of knowledge, such as genes, diseases, chemical substances, organisms, genomics, phenotypes, and more (Unni et al., 2022).

The Biolink Model is designed with an object-oriented classification and graph-oriented features, and was developed collaboratively by researchers in the biomedical informatics field, ensuring that it complies with the needs of researchers and aligns with ongoing scientific projects and ecosystems. Similar to ontologies, it includes class hierarchies, allowing the inheritance of properties and relationships. Biolink is used by large-scale informatics projects such as the Monarch Initiative providing a structured representation of biological entities and their relationships (Putman et al., 2024).

[4] https://www.w3.org/TR/skos-primer/

3.10. OMOP-CDM

The Observational Medical Outcomes Partnership Common Data Model[5] (OMOP-CDM) is a healthcare data model developed by the Observational Health Data Sciences and Informatics (OHDSI) community. OMOP-CDM facilitates standardization by defining a common tabular structure for healthcare data. Each table describes a specific domain of clinical observations, such as patient demographics, condition occurrences, drug exposure or clinical measurements (I et al., 2021). OMOP-CDM uses a Web-based lookup service from a standard vocabulary, called ATHENA, as its source of annotations.

Key features of the OMOP-CDM include its integration with a plethora of analytical tools developed by OHDSI community, enabling the common environment for users to perform reproducible analysis.

3.11. Shape Expressions (ShEx)

Shape Expressions[6] (ShEx) is a Web standard for describing and validating RDF graph structures. ShEx allows the definition of properties, cardinality constraints, data types, and value ranges that are expected to appear for a specific class of data represented in an RDF graph. ShEx validation assists in achieving data consistency, making it easier to integrate across datasets. The ShEx validation tool used in this thesis is RDFShape[7], though the selection of this tool was arbitrary.

3.12. RDF Mapping Language and YARRRML

RDF Mapping Language[8] (RML), is a language for directing the transformation of heterogeneous and diverse data sources, like relational data, into RDF graphs. This technology is based on R2RML[9] (RDB to RDF Mapping Language), a W3C standard designed for mapping relational data to RDF graphs. RML extends the capabilities of R2RML by supporting a wider range of input data formats, including XML, JSON, or CSV. RML mappings specify how to extract data from these

[5] https://ohdsi.github.io/CommonDataModel/
[6] http://shex.io/
[7] https://rdfshape.weso.es/
[8] https://rml.io/specs/rml/
[9] https://www.w3.org/TR/r2rml/

sources and transform it into RDF triples. RML mappings are defined using a declarative syntax, however, RML documents themselves are not intended to be human-readable, and are therefore difficult to maintain.

YARRRML[10] is a declarative language for creating Linked Data mapping rules. This language can be used for authoring RML rules in a more human-friendly manner. YARRRML is expressed in YAML syntax, which is more readable than the RDF syntax used by RML.

3.13. REST APIs

An Application Programming Interface (API) is a set of rules, protocols, and tools that allows different software applications to communicate with each other. APIs define the methods and data formats that applications can use to request and exchange information.

REST, or Representational State Transfer, is an architectural style for designing networked applications. REST architectures are distinguished by having few operations, but myriad uniquely-identified resources. In the context of the Web, HTTP methods define the operations, and these operate on data objects identified by URLs. Further, the resource identified by the URL may have multiple representations, for instance, JSON or XML.

REST APIs on the Web use the HTTP protocol to pass messages between a client and a server. The HTTP methods used by REST APIs in the bioinformatics space are generally limited to the core five: GET, PUT, POST, HEAD, DELETE. REST tightly defines what the behavior of each of these methods should be, as follows:

1. GET: Retrieve a resource.
2. POST: Create a new resource that is a subordinate of an existing resource
3. PUT: Create a new resource, or entirely an existing resource.
4. HEAD: Retrieve metadata about a resource.
5. DELETE: Remove a resource

[10] https://rml.io/yarrrml/spec/

Resources can be represented in various formats (the "Representational" part of the REST acronym), for instance: JSON, XML, or RDF. The client can use "content negotiation" via HTTP headers to indicate to the server its preferred representation for a given resource.

4. Materials

In this section, the tools, vocabularies, ontologies, and documentation provided by third-parties and used during my thesis are described. The tools I created during this research project are described in the subsequent chapter of Methods and Results.

4.1. EC RD Platform Common Data Elements

The European Platform on Rare Disease (EU RD Platform) Registration has defined a set of 16 Common Data Elements[11] (CDEs) for RD registration, and their allowed values. These are detailed in Table 1.

Element ID	Name	Accepted value
1.1	Pseudonym	String
2.1	Date of birth	dd/mm/yyyy
2.2	Sex	Female, Male Undeterminated, Foetus (Unknown)
3.1	Patient status	Alive, Dead, Lost in Follow-up, Opted-out
3.2	Date of death	dd/mm/yyyy
4.1	First contact with specialized centre	dd/mm/yyyy
5.1	Age of onset	Antenatal, At Birth, Date (dd/mm/yyyy), Undeterminated
5.2	Age at diagnosis	Antenatal, At Birth, Date (dd/mm/yyyy), Undeterminated

[11] https://eu-rd-platform.jrc.ec.europa.eu/sites/default/files/CDS/EU_RD_Platform_CDS_Final.pdf

6.1	Diagnosis of the rare disease	ORPHA Code, Alpha Code, ICD9/10 Code, ICD9-CM Code
6.2	Genetic diagnosis	Human Genome Variant Sequence (HGVS), HUGO Gene Nomenclature Committee (HGNC), Online Mendelian Inheritance in Man (OMIM) Codes
6.3	Undiagnosed case	Human Phenotype Ontology code and/or HGVS Code related to the inability to diagnose.
7.1	Agreement to be contracted for research purposes	Yes/No
7.2	Consent to reuse data	Yes/No
7.3	Biological sample	Yes/No
7.4	Biobank identifier	URL/No
8.1	Disability Classification via International Classification of Functioning and Disability (ICF)	Score

Table 1. Common Data Elements, grouped by domain of knowledge, including the permitted value of each.

4.2. SDM-RDFizer and RML Mapper

Both SDM-RDFizer (Iglesias et al., 2020) and RML Mapper[12] belong to a group of tools called RDFizers. RDFizers transform various data formats, such as CSV or JSON into RDF using RML mappings to define how the data source should be transformed into triples or quads.

4.2.1. SDM-RDFizer

SDM-RDFizer is a RDFizer written in Python capable of transforming CSV, JSON, XML and relational databases into RDF. SDM-RDFizer is designed for high performance transformation making it suitable for processing datasets quickly.

[12] https://github.com/RMLio/rmlmapper-java

4.2.2. RML Mapper

RML Mapper is an RDFizer written in Java, capable of transforming several data sources like CSV, JSON and XML formats into RDF triples and quads. RML Mapper is capable of defining custom functions for conditional operations, expanding the variety of ways to create RML mapping. RML Mapper is part of a larger open-source ecosystem from a team of researchers at Ghent University, including tools such as the YARRRML Matey[13] parser, an efficient YARRRML to RML parser.

4.3. GraphDB

Developed by Ontotext, GraphDB[14] is a high-performance triplestore designed for storing and querying RDF data. Triplestores are database-like repositories that, unlike common relational databases, do not organize data into tables. GraphDB provides robust support for RDF data management and querying. It supports OWL and some limited logical inference, various RDF syntaxes (N-quads, Turtle, N-Triples) and the SPARQL query language. GraphDB provides a free version, supported by developers, and this is the version used in this thesis, via its public docker image.

4.4. Docker and Docker compose

Docker is a platform for distributing and running applications that do not require the user to install the native application software on their system. Docker allows developers to package an application and all its required dependencies (language interpreters, modules, packages, databases) into a standardized unit called an "image", facilitating consistent behavior across any machine that executes that image. The most significant components of the Docker infrastructure include:

- Docker images: Immutable snapshots of an application and its dependencies. Images are built from a set of instructions defined in a Dockerfile, and ensure that all users receive the correct versions of all dependencies.

[13] https://rml.io/yarrrml/matey/

[14] https://graphdb.ontotext.com/

- Docker containers: Portable and isolated environments created by "running" Docker images, shown in Figure 5.

- Docker volumes: Persistent storage for containers, allowing data to be stored and shared independently of the container lifecycle.

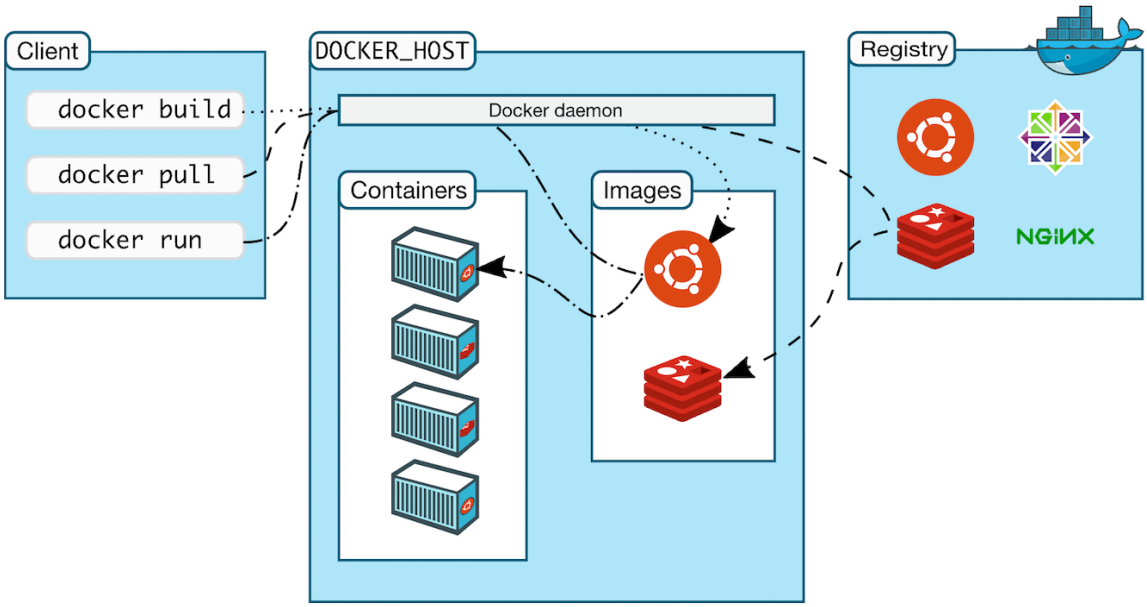


Figure 5: Overall Docker architecture.

Docker uses a script with a set of instructions for building Docker images. Executing this script causes Docker images to be packaged in a repeatable and consistent manner. Running this image results in the creation of a Docker Container, that is the actionable component of the Docker ecosystem. Containers may persist, holding their state between sessions, or they may be deleted between sessions, thus reverting to their base state.

Docker Compose is a tool designed to manage one or multiple Docker containers that are intended to work together. This orchestration is controlled by a YAML file, used to configure the Containers, inter-Container networks, and shared Volumes.

4.5. FAIR Data Point (FDP)

A FAIR Data Point (FDP) is a FAIR-compliant set of publishing standards for metadata about catalogs and datasets, compliant with existing standards, such as the Data Catalog Vocabulary[15] (DCAT) and the Linked Data Platform[16] (LDP).

The FDP specification also defines a REST API for creating/updating DCAT records. The reference implementation of the FDP specification includes an FDP Client that allows the creation and editing of metadata. The FDP Reference Implementation uses the Shapes Constraint Language[17] (SHACL) for defining Web interfaces and for data type validation.

4.6. FAIR-in-a-box (FiaB)

FAIR-in-a-box[18] is a bash script that acts as an installer and configuration bootstrapper for a set of Docker images that can be combined, via docker-compose, into a complete RDF-based ETL workflow. The components that are installed by FiaB include:

- Tools for RDF serialization, including RDFizer and YARRRML parser.
- GraphDB interface for data storage
- FAIR Data Point (FDP) Client and Server.

4.7. OHDSI interfaces

OHDSI is an international collaborative effort that aims to create a comprehensive and open science network in the health sciences. To do so, OHDSI designs and implements advanced data analytics and an extensive set of tools and resources to support data standardization (Hripcsak et al., 2015; I et al., 2021) . The cornerstone of these goals is the creation of OMOP-CDM, which allows harmonization across distinct healthcare data sources into a common format. Some of the key componentes OHDSI interface includes and are used in this thesis are:

[15] https://www.w3.org/TR/vocab-dcat-2/
[16] https://www.w3.org/TR/ldp/
[17] https://www.w3.org/TR/shacl/
[18] https://github.com/ejp-rd-vp/FiaB

4.7.1. ATHENA

ATHENA is a comprehensive Web-based lookup service developed by the OHDSI community. ATHENA was introduced in 2015, serving as a Web lookup service for exploring standardized vocabularies (Reich et al., 2024). This vocabulary includes healthcare ontologies and medical terminologies such as ICD-9, ICD-10, SNOMED, LOINC, and RxNorm. ATHENA facilitates and supports the integration of the diverse healthcare data sources into the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM).

4.7.2. Data Quality Dashboard

The Data Quality Dashboard[19] (DQD) is an interface developed by the OHDSI community to expose and evaluate observational data quality within the OMOP-CDM. The DQD evaluates various data quality dimensions such as completeness, conformance, consistency, and plausibility, across a given dataset, creating a resulting report of the data compatibility with OMOP-CDM.

[19] https://ohdsi.github.io/DataQualityDashboard/

5. Methods and Results

5.1. Is it possible to create a FAIR-compliant data model for the domain of rare diseases?

5.1.1. Specifying the necessary components of the model

In the following paragraphs, **bold face** is used to indicate a specific type of entity or process that will need to be represented as an RDF resource.

Typically, each datum in a clinical research repository is derived from an assessment **process**, where the **subject**, typically a patient, comes to a specialized center where a clinician or other designated specialist observes and/or measures the condition and personal **attributes** this patient manifests. Both the patient and the qualified specialist therefore participate in this clinical assessment process.

The exact clinical process is dependent on the **type of observation or measurement** that occurs during that process. For example, the diagnostic process for a condition differs from that of assessing body mass index or blood pressure, even if these procedures occur during the same clinical encounter.

Every clinical assessment has a **measurement output**. In the case of quantitative observations, this measurement output will generally consist of a **value** and its **unit of measurement.** For instance, the blood pressure measurement process has an output blood pressure measurement value, further defined by the unit of millimeters of mercury (mmHg). Moreover, each of these measurement outputs represent an **attribute** of the patient at the time the measurement was made. For example, during a body measurement assessment process, the measurement value is associated with a personal attribute such as height. For qualitative observations, the output is generally a judgement (formally called an **"information content entity"**) with a meaning that can often be captured using a controlled vocabulary or ontology term.

The subject, particularly the patient, is identified by a **patient identifier** in research registries. We find, however, that this is an imprecise way to envision the

role of an identifier; rather than identifying a person, they identify a patient - that is, the **role** of "patient" that the person plays during a healthcare activity. Thus, a person may have multiple identifiers, each denoting a distinct role that person plays in various activities. The same person will be associated with, for example, a driver's license identifier, which is meaningful in the context of their driving. Therefore, the Role entity serves as a connection between the identifier and the process within which the role is realized, such as the patient role in this case. This patient identifier is generally an alphanumeric patient unique identifier with a local context, meaning that it generally will have no meaning outside of the context of that data registry.

Through exploration of the context associated with numerous clinical observations captured by the EJP-RD registries, we noted that every observation consisted of this same set of concept/relations that describe the clinical encounter:

- Subject
- Role
- Process
- Attribute
- Process Type
- Identifier
- Measurement
- Value/Unit

While these concepts were seldom explicitly noted in the source clinical data, they can be inferred to exist. Moreover, besides the measurement itself, these concepts are related to how the datum was measured, and what entities were involved in the measurement - in this regard, they are a form of provenance metadata.

From this starting point, an informal map of the entities and entity-relations was derived that appeared capable of representing all of the concrete observations within the clinical registry datasets.

5.1.2. Selecting upper ontologies for the model

To formalize the crude concept map described above, ontological modeling activities were undertaken by a core group of experts from the EJP-RD project, to

achieve agreement on the ontologies used and the design pattern for this semantic data model.

There are multiple ontologies, called "upper ontologies" that define basic features of reality - time, space, and qualities - and how they relate to one another. Ontologies like Basic Formal Ontology (BFO) (Otte et al., 2022) and Semanticscience Integrated Ontology (SIO) present a broad selection of ontological terms for conceptually representing this upper-domain of lifescience knowledge. Most of these ontologies contain broad relationships between these entities, each one with a similar lexical definition and hierarchy under their schema. Moreover, anticipating likely future expansion for more data elements, the model should be designed in a forward-looking manner.

SIO was selected as the upper ontologies for our semantic model due to its comprehensive biomedical description, facilitating life science knowledge discovery and interoperability. Moreover, SIO is capable of describing not only the provenance of the clinical processes, but also the essential healthcare concepts and relationships, and how they can be interpreted for scholarly inquiry. SIO has a well-documented set of design patterns[20] that assist in the creation of logically consistent data models.

5.1.3. Selecting domain-specific ontologies for the data model

Beyond the upper ontologies that define core concepts and their relationships, we also endeavoured to identify domain-specific concepts in life science-oriented ontologies. The purpose of this identification was to find a suitable ontology or ontologies that could represent the biomedical concepts required by the clinical data elements we have selected for semantic transformation. National Center for Biomedical Thesaurus (NCIT) (Ceusters et al., 2018; Fragoso et al., 2004) , Provenance ontology (PROV-O) (Lebo et al., 2013) and Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) (El-Sappagh et al., 2018) are some domain-specific examples of these ontologies.

[20] https://github.com/MaastrichtU-IDS/semanticscience/wiki/Design-Patterns

Global guidelines that constrained the selection of domain-specific ontologies were discussed. First, the use of the least number of ontologies possible was one the criteria for the creation of this semantic model. Another important consideration is the absence of restrictive licenses that could create constraints in certain countries. For instance, SNOMED CT is a well-established ontology in the healthcare community; however, the SNOMED CT license is restrictive in certain countries outside the EU, limiting its inclusion in the semantic data models of rare disease patient data that should be used within international collaborations.

The PROV-O aims at supporting process representation that generates or manipulates these clinical entities. PROV-O is designed to capture the provenance of a process - that is, the participants, the process itself, and its inputs and outputs. Thus, while it could capture information related to the act of collecting a clinical observation, it is not capable of representing the quantitative or qualitative details of that observation, nor how they relate to the actors involved in the data gathering process.

Alternatively, OBO Foundry ontologies, especially NCIT, provide a nearly complete representation of the biomedical concepts required by the clinical data elements. OBO Foundry ontologies were chosen for describing the domain-specific aspects of the core structure for given instances of the model - i.e., what kind of role? What kind of attribute?

Other annotation vocabularies for biomedicine and molecular biology were identified and selected for precisely specifying the clinical diagnosis and pathogenic genotype-phenotype annotations.

- Human Phenotype Ontology (HPO) for phenotypic abnormalities in human diseases (Köhler et al., 2021).
- Orphanet Rare Disease Ontology (ORDO) for rare diseases derived from the Orphanet database (Vasant et al., 2014).
- Human Genome Variation Society (HGVS) (Dunnen et al., 2016) and Online Mendelian Inheritance in Man (OMIM) (Hamosh et al., 2000) for describing DNA, RNA, and protein sequence variants. Public Databases for genotype information, such as ClinVar, uses HGVS and OMIM as an standardized annotation for representing gene variations.

5.1.4. Data modelling with SIO and OBO

SIO design patterns include relationships that we can use to, for example, describe how a patient identifier denotes a patient role, and how a person takes-on that role as a result of their participation in a particular clinical process. From this clinical process, how an output is generated, holding the measurement or observation value related to an attribute of the person playing the patient role. Tables 2 and 3 specify the entities and relationships defined in SIO that were selected to be used in our core data model.

SIO annotation code	Label	Definition
SIO_000498	person	A person is an object that has certain capacities or attributes constituting personhood.
SIO_000115	identifier	An identifier is a label that specifically refers to (identifies) an entity (instance/type).
SIO_000016	role	A role is a realizable entity that describes behaviours, rights and obligations of an entity in some particular circumstance.
SIO_000006	process	A process is an entity that is identifiable only through the unfolding of time, has temporal parts, and unless otherwise specified/predicted, cannot be identified from any instant of time in which it exists.
SIO_000015	information content entity	An object that requires some background knowledge or procedure to correctly interpret.
SIO_000614	attribute	An attribute is a characteristic of some entity.

Table 2: core SIO classes for common data elements definition. The SIO annotation code column uses only the unique portion of the concept URI.

SIO annotation code	Label	Definition
SIO_000671	has identifier	a relation between an entity and an identifier.
SIO_000228	has role	is a relation between an entity and a role that it bears.

SIO_000020	denotes	is a relation between an entity and what it is a sign or indication of, or what specifically means.
SIO_000356	is realized in	No label included in the ontology.
SIO_000229	has output	is a relation between a process and an entity, where the entity is present at the end of the process.
SIO_000628	refers to	is a relation between one entity and the entity that it makes reference to.
SIO_000008	has attribute	is a relation that associates a entity with an attribute where an attribute is an intrinsic characteristic such as either directly or indirectly through generalization of entities of the same type.a quality, capability, disposition, function, or is an externally derived attribute determined from some descriptor
SIO_000300	has value	A relation between an informational entity and its actual value (numeric, date, text, etc).

Table 3: core SIO properties for common data elements definition. The SIO annotation code column uses only the unique portion of the concept URI.

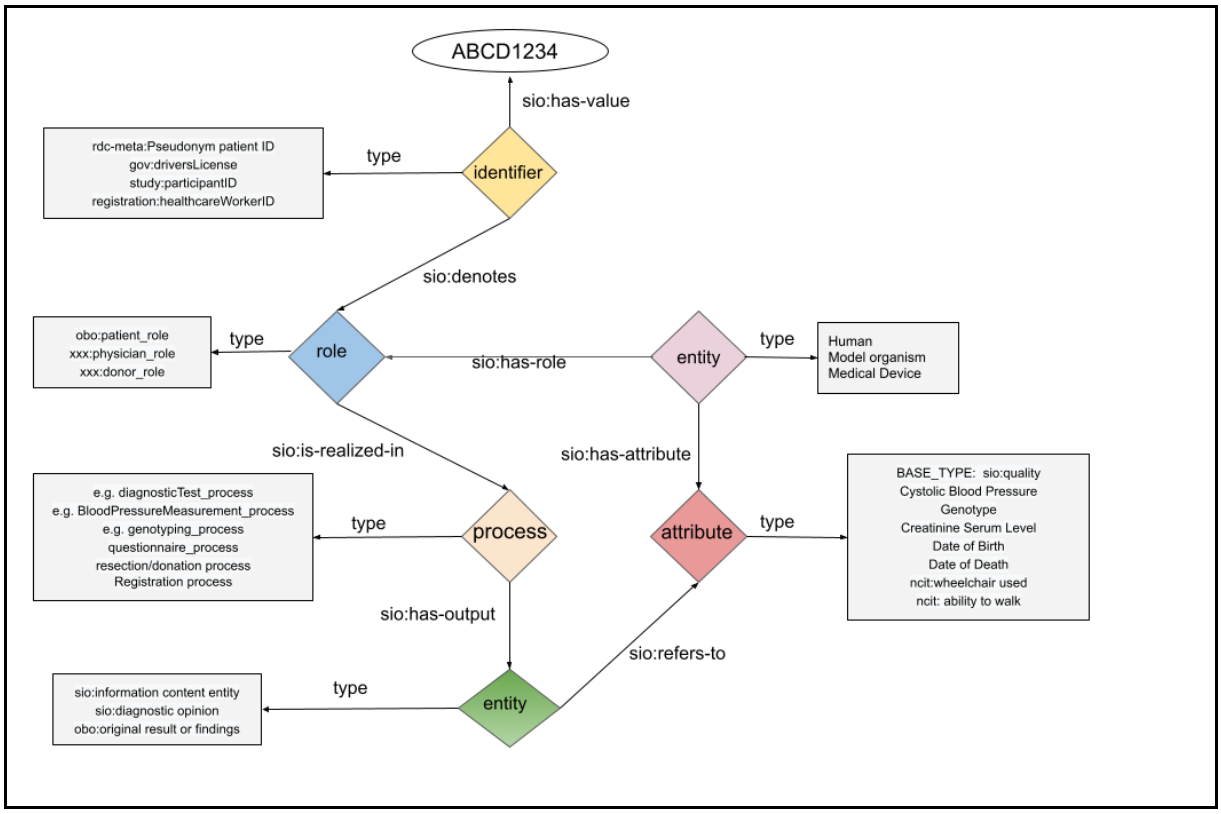


Figure 6. SIO-based core metadata structure of the CDE-SM Model.

These combined ontological types result in the model shown in Figure 6, named the Common Data Element Semantic Model (CDE-SM) (Kaliyaperumal et al., 2022). Following the creation of the CDE-SM, we then designed a specific model for each of the 16 CDEs in the RD Platform, shown in Table 1. We noted that the CDEs are not uniformly independent from one another, and thus we sometimes merged CDEs in our CDE-SM models. For example, the Patient Status CDE-SM model includes a patient status of being "dead", which is associated with a date of death; however, these two CDEs are defined independently in the RD Registration Elements (3.1 and 3.2 in Table 1). Similarly, the CDE-SM consent model can be reused for diverse types of consent (e.g., consent for contact, consent for data reuse), which are separate in the RD Registration Elements. Finally, the Pseudonym data element is a part of every other model. The resulting list of CDE-SM models, mirroring the RD Registration CDEs, is shown in Table 4, and its figure in Annexes A.

Data element name	**Purpose**
Disease Progression	A "container" node to group together all other CDEs that refer to the same diagnosis. For example, the "age of diagnosis" CDE is related to a specific rare disease via traversal into the "disease progression" container, and then traversal into the "diagnosis" CDE that is also connected to "disease progression"
Care Pathway	Captures the date of first contact with the specialist healthcare system; is connected to "disease progression"
Diagnosis	Captures the final disease diagnosis using ORPHA codes; is connected to "disease progression"
Disease History	Captures age at first symptoms and age at diagnosis; is connected to "disease progression"
Genetic Diagnosis	Captures the sequence variant(s) found in this patient, using a variety of different coding systems; is connected to "disease progression"
Patient Consent	Captures the consent of the patient over several axes (e.g., consent for contact, consent for data reuse). Provides a reference to the signed consent form, as well as an input reference to the (blank) consent template.

Patent Status	Captures the current status of the patient, and their date of death if the patient is deceased
Personal Information	Captures (superficial) personal information such as birth date and sex (there are ongoing debates in the EJP modelling group as to whether this should be converted to an age, or an age-range, for improved privacy)
Phenotyping	Captures the phenotypes of the patient, using Human Phenotype Ontology terms
Disability	Captures the score for a disability test. The specific test administered is indicated as one of the child nodes of obo:NCIT_C20993 (Clinical or Research Assessment Tool), and thus this CDE model is broadly useful for many disorders.
Undiagnosed	Captures the case where a patient has phenotypic anomalies, and an identified sequence variant, but for some reason has not been definitively diagnosed.

Table 4: Models created to represent the common data elements.

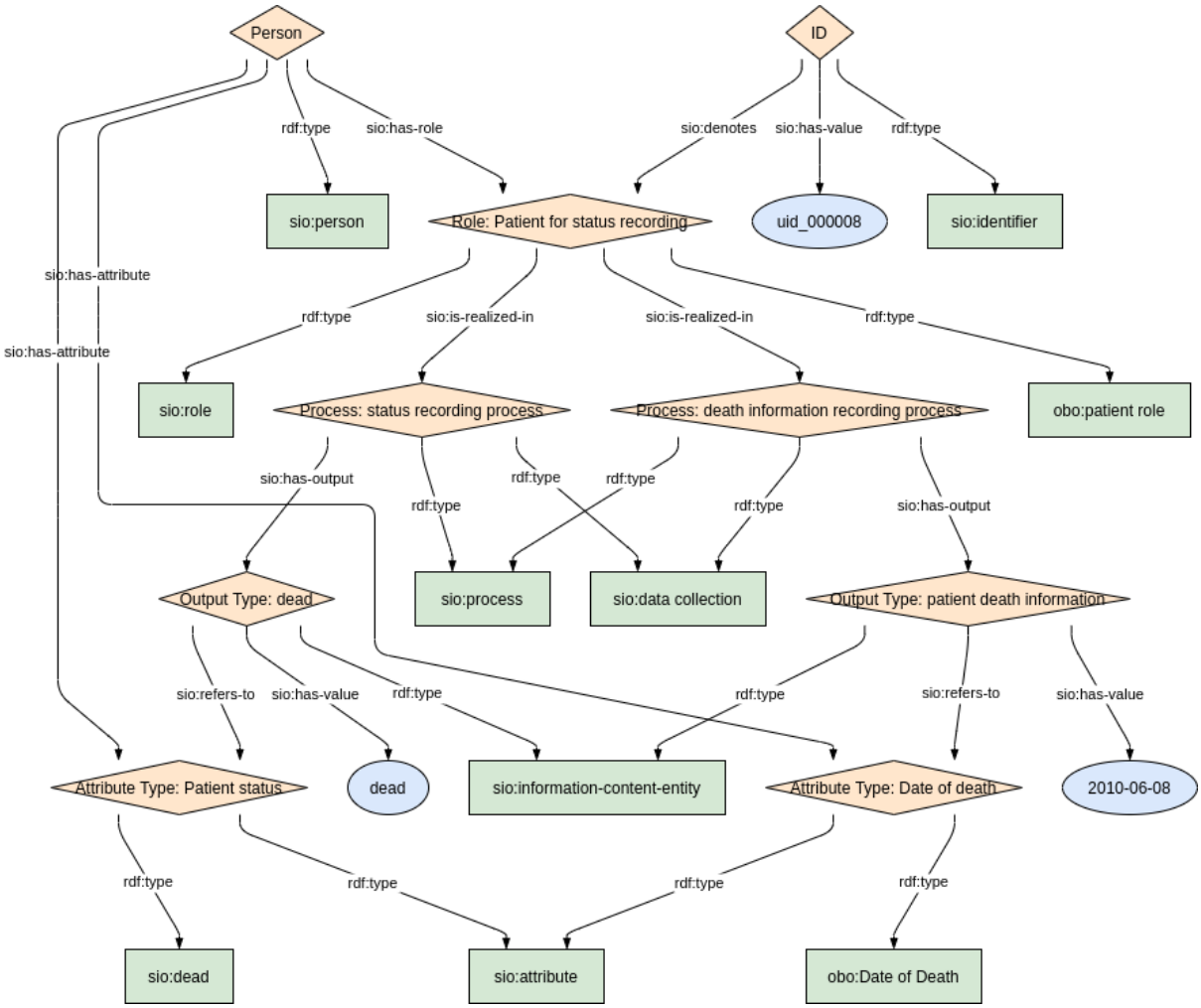


Figure 7: Exemplar representation of patient status information as an instance of the CDE-SM models participation status and date of death.

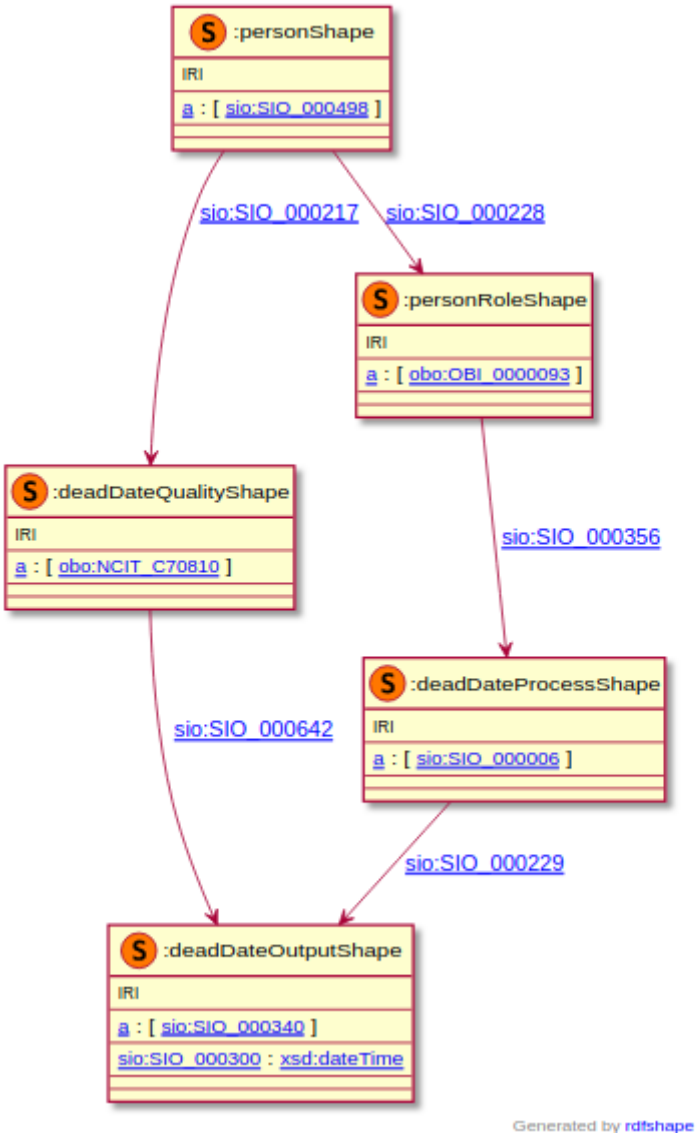


Figure 8: ShEx representation of death date information.

In Figure 7, the representation of one of these data elements is visualized. Each type of semantic structure is represented by a different color and shape; orange diamonds represent instances; green rectangles are classes and blue rectangles represent literal data values. All these entities are interlinked by properties, represented as arrows.

Every data element model was mirrored by the creation of a ShEx file that can be used for data validation. These ShEx files describe both the classes contained by each instance, and their valid class-relationships. In case of data elements in which all value options are defined by a controlled vocabulary, this is indicated in the ShEx file, for instance, by providing every option for "Sex" (Female, Male, Unknown or Undetermined) according to its ontological URI. Figure 8 presents the

visual representation of the ShEx corresponding to the death date data element. All such data elements were visually represented by the use of RDFShape, a Web tool for several implementations related with RDF-based artifacts, such as RDF, XML and RDF valuation like ShEx and SHACL.

5.2. Can we define a workflow, and its associated tooling, that enables non-FAIR experts to transform their data into these models?

5.2.1. The use of CSV and YARRRML for semantic data serialization.

With the large number of registries that participate in the EJP-RD, it is not possible to create a novel FAIRification solution for each participant. Moreover, the data is in many cases sufficiently sensitive that the project's FAIR experts were not allowed to manipulate it (or even see it), so it was not going to be plausible for the FAIR experts to work with the data themselves. As such, it was clear that a near-fully automated system was needed that could be shared and implemented at all participating sites.

Comma-Separated Value (CSV) is a data format that, while quite limited in expressivity, can be generated from most starting formats/technologies (e.g. RDBMS, Excel, JSON, etc.), thus it could be generated easily by all participants. CSV, then, was adopted as the *lingua franca* for the EJP FAIRification efforts, and we subsequently focused on tools and technologies capable of transforming CSV into RDF. YARRRML, as a declarative syntax for RML rules, was adopted as a templating language that would be easier to maintain and debug versus the more complex RML documents it represents, and thus YARRRML was selected as the primary means of sharing the CDE-SM templates, with the YARRRML to RML conversion happening at the time of transformation. RML supports CSV as a data source for its mapping functionality.

5.2.2. YARRRML builder for creating templates

Although YARRRML is a more human-friendly language/syntax compared to RML, the creation and maintenance of YARRRML templates nevertheless make scalability a challenge. To mitigate this time-consuming issue, we created a Python-based YARRRML builder that uses Python objects to represent triple-fragment templates, making it straightforward to write YARRRML templates in code.The use of this YARRRML builder benefits through:

1. The creation of a more human-friendly syntax than YARRRML, where the code presents only a Python List Object, making it easier to explore and validate.

2. Custom functions for managing the data source path and base IRI prefix used.

3. YARRRML builder reuses the same Python object to create both schema and query representations, such as ShEx or SPARQL, allowing it to easily keep all of these representations synchronized over the entire project.

These 3 aspects reduce the manual curation of the templates, increase automatization, and enhance reusability.

5.2.3. Definition of CSV templates

Each CSV column header is a reference into the YARRRML/RML template. Thus, these headers must be pre-defined for any CSV file that will be passed through our ETL pipeline. It is also necessary to consider what values will be allowed in each CSV template column.

For globally unique identifiers, or references to formal vocabularies or ontologies URLs must be provided. YARRRML/RML facilitates the definition of prefixes, so ontological references can be added in their abbreviated form using a namespace.

Each data value must adhere to a specific datatype. YARRRML/RML allow defining XML Schema Definition (XSD) datatypes in the template, so each value can be explicitly typed (e.g., xsd:integer, xsd:float, xsd:date).

For ensuring date information to be consistently formatted, the use of ISO-8601, an international standard for date and datetime notation (Briney, 2018) - and the standard for RDF – is enforced.

Based on these considerations, documentation[21] for creating CSV templates was written to guide data sources in properly constructing these intermediate data representations, and limiting the possible values.

5.2.4. How to automatically build RDF data entities that don't "collide"

In RDF, every entity is identified by a URI. It is a feature of RDF that, when the same URI is encountered, it is (by definition) referring to the same entity. Thus, during an automated RDF transformation, the construction of unique URIs for every entity is a critical requirement (and conversely, the reuse of an entity's URI every time it is referenced is also a requirement). We refer to the reuse of a URI for distinct entities to be a "collision", and it irreparably corrupts the data. Blank nodes (i.e. RDF nodes that do not have a defined URI, but where the URI is automatically generated by the software that is consuming it) are commonly used for this purpose; however, this does not result in URIs that can be externally referenced, and eliminates the possibility of reusing the same URI for the same entity.

Generating unique URIs in our transformation was achieved by the use of a unique timestamp. These timestamps are added in the templated URI structure in the YARRRML file, thus ensuring that all URIs in the output RDF data are unique when they should be unique, and are shared when they should be shared.

We built a custom script for this timestamp generation. This script is executed before the RDF serialization, creating a column in the CSV template containing a unique identifier for each observation.

[21] https://github.com/ejp-rd-vp/CDE-semantic-model-implementations/tree/master/CDE_version_1.0.0/docs

5.2.5. RDF serialization via CSV and YARRRML templates

The use of a YARRRML/RML template, and an appropriate data source (CSV for this case) are the inputs to RDF transforming software commonly called "RDFizers". Several RDFizers were tested to identify their computational efficiency and speed.

RML-Mapper and SDM-RDFizer were selected as suitable solutions. Although RML-Mapper allows custom functions and a rich set of features, including conditions driven by logical operations, SDM-RDFizer is computationally faster, and thus was the selected RDFizing technology for this phase of the project.

SDM-RDFizer was "wrapped" as a RESTful Web Service, and a Docker image of this service was built to facilitate placing it into a larger ETL workflow. Association between the request for data transformation received by the Docker container, and the correct YARRRML model/CSV file, was regulated by a "tag" passed as an argument into the request call; all CSV and YARRRML files therefore had strict naming conventions taking advantage of this tag.

5.2.6. RDF validation and storage

After RDF generation of each data element, it was tested to ensure the resulting data adhered to the expected model. Each of the CDE-SM models has an associated ShEx file that is used for this validation.

To perform the evaluation, the RDFShape Web tool was selected. This interface, also used to generate the visual representation of the ShEx files (as it's shown in Figure 8), performs RDF data validation against ShEx and other languages such as SHACL. Validation requires both ShEx and RDF files, along with a ShapeMap definition. ShapeMap describes a union node between the RDF sample and the ShEx file.

Conformant RDF data was then stored in GraphDB through an automated upload process using the GraphDB REST API, and a "daemon" component of the FiaB pipeline.

5.3. Is CDE-SM flexible enough to represent data elements beyond the CDEs?

5.3.1. Mock data synthesis

ERKreg (Bassanese et al., 2021), a participant registry in EJP-RD, contains clinical information of vascular rare disease patients from several countries in Europe. ERKreg allowed us to use a subset of anonymous patient data from their registries for open testing purposes and benchmarking. We have used this subset of patient data entries to mechanically generate a larger synthetic yet representative patient sample for mapping and testing the capabilities of CDE-SM to be FAIR-transformed and represented by RDF. This initial subset contained 20 patients, described with multiple data entries dedicated to multiple domains of clinical knowledge:

- Personal information, such as patient sex, participation status.
- 12 different vascular rare disease diagnosis codes, annotated by Orphanet.
- Molecular analysis, both genetic and laboratory measurement.
- Symptoms, in the form of abnormal phenotypes (e.g.: Vasculitis).
- Treatment related descriptions and medications, annotated by Anatomical Therapeutic Chemical (ATC)
- Policy for patient consensual agreement

Temporal information was included in the patient data from ERKreg, keeping patient information fully anonymous. Temporal information was created by GPT-like technologies, creating a table that contained several dates that corresponded with the human timeline, sorted from first to last. Then, ERKreg data elements were injected with the synthetic temporal information to ensure that:

- The data included a realistic timeline of events where some patient data description occurred before others, for instance: date of birth before symptoms, or molecular analysis before diagnosis.
- We avoided exposing patient data (anonymous or not) via a GPT prompt.

For data elements in the EC RD Platform CDE list that were not present in this subset, such as Biobank availability and disability assessment, this information

was synthetically added by manual curation so that every required data element was present. In total, 100 synthetic patients were created.

In order to simulate real patient registry entries, not every data element was described in all the subjects. Even so, all CDEs were described in the patient data sample synthesized. This, therefore, can now be used as a reference dataset for future model validation by both EJP-RD and other clinically-oriented projects.

5.3.2. Identifying limitations and inconsistencies

Limitations of CDE-SM	Proposed solutions
Inconsistency in aggregated data elements	Individually representation of every data element
Inconsistency in ontological terms	Standardization of the ontological classes
Lack of proper longitudinal data definition	Semantic metadata layer for encounters and temporal description
No place to capture important provenance information such as protocols/workflows	Expansion of core set of data elements

Table 5: Summary of limitations and proposed solutions for the semantic data model.

Through several years of use, the CDE-SM model revealed itself to be lacking in several important ways (shown in Table 5) that made it increasingly difficult to expand into new data types, and to map into other clinical data standards. In addition, there was a need to specify inputs into processes, and to better capture provenance information by attaching protocols to the processes. As such, we began exploring a new model that addressed these limitations.

5.3.3. Limitations and inconsistencies resulting from CDE aggregation

CDE-SM adopted the high-level data elements defined by the RD Platform, but often needed to aggregate these to achieve semantic consistency. As a consequence, registries did not always have all of the individual subcomponents to fulfill a given CDE-SM model, leading to incomplete FAIR-transformed data or data loss.

Moreover, temporal information was sometimes included in the model, and sometimes not, depending on the model representation and the ability of data registries to provide it.
Furthermore, not every data element used the entire core structure defined by SIO, not facilitating the creation of a common SPARQL for every data element.

Each CSV files was associated with a distinct YARRRML template, causing concerns about the maintenance and sustainability of the existing templates, and making the modelling of new data elements quite tedious, despite the availability of the YARRRML Builder tool, and especially as the number of data templates was growing.

To address these problems, we undertook the following steps, which will be described in more detail in subsequent paragraphs:

1. Spliting previously aggregated data elements into distinct representations. Each of these highly granular data elements were then documented and represented by a single model mitigating incomplete model filling during transformation.

2. Unifying the YARRRML into a single template that represents all data elements - both CDEs and the expanded set.

3. Concomitant unification of the CSV template, which is documented with the instructions for which CSV columns are required to be populated for each data type, and other details such as how to format dates.

Unification of the YARRRML templates necessitated a new feature not required by the CDE-SM templates - that is, conditional functions that determine if an element is going to be used in the transformation. This need arises in cases where not all components in the core semantic model are needed to achieve a complete representation of the data, and thus these model components can (and should) be ignored during transformation. SDM-RDFizer is not capable of managing functions of any kind, and so a new YARRRML parser and RDFizer were needed. SDM-RDFizer was therefore replaced with RMLMapper, another of the options identified as a suitable RDFizer in the earlier phase of the project. Importantly, we had also discovered that the efficiency of SDM-RDFizer was achieved by

holding all data in memory, and this resulted in "silent" data loss when large datasets were transformed under a single execution. Thus, this component had to be replaced for data quality reasons, despite any loss in efficiency.

5.3.4. Limitations and inconsistencies with ontological terms

In CDE-SM, the number of ontological terms used to define every model node was arbitrary, and had no grounded rationale or justification. In some cases, ontological terms were redundant, for example, having two or more terms from the same ontology, one of them being a parent concept. Moreover, some of these cases were not parent-child concepts but disjointed concepts from the same ontology, leading to semantic inconsistency.

This redundancy required a global revision in the number and ontologies used for defining the new models. The new requirements were that every sub-component of the data model should be typed through a single SIO upper-level ontological class, and a single domain-specific one from OBO Foundry (mostly described by NCIT). As an alternative to OBO, any of the ontologies used for diagnosis or genotype-phenotype annotations, such as Orphanet or OMIM, were allowed when necessary.

5.3.5. Limitations and inconsistencies with longitudinal data element representation

Patient clinical records are complex - even more so when considered longitudinally. A single patient may have multiple encounters with the healthcare system, for multiple reasons, for distinct disorders. The separation between these is not necessarily distinct in time or space, and thus you may have overlapping symptoms that are associated with distinct diagnoses happening at the same time. The CDE-SM model considered every clinical observation as an atomic unit, unrelated to any other observation. Thus, symptoms couldn't be related to diagnoses, and treatments could not be related to symptoms (or the resolution of symptoms). Together, this makes the dataset unsuitable for research purposes.

While the EC RD Registry CDEs had a "Disease progression" data element, it was entirely insufficient to capture these inter-relations in the patient's clinical record. Though the CDE-SM model for this element was intended to allow the creation of a progression of patient observations, our implementation was not capable of

interlinking models that were created in previous serializations, since every data element was made unique by the inclusion of a timestamp.

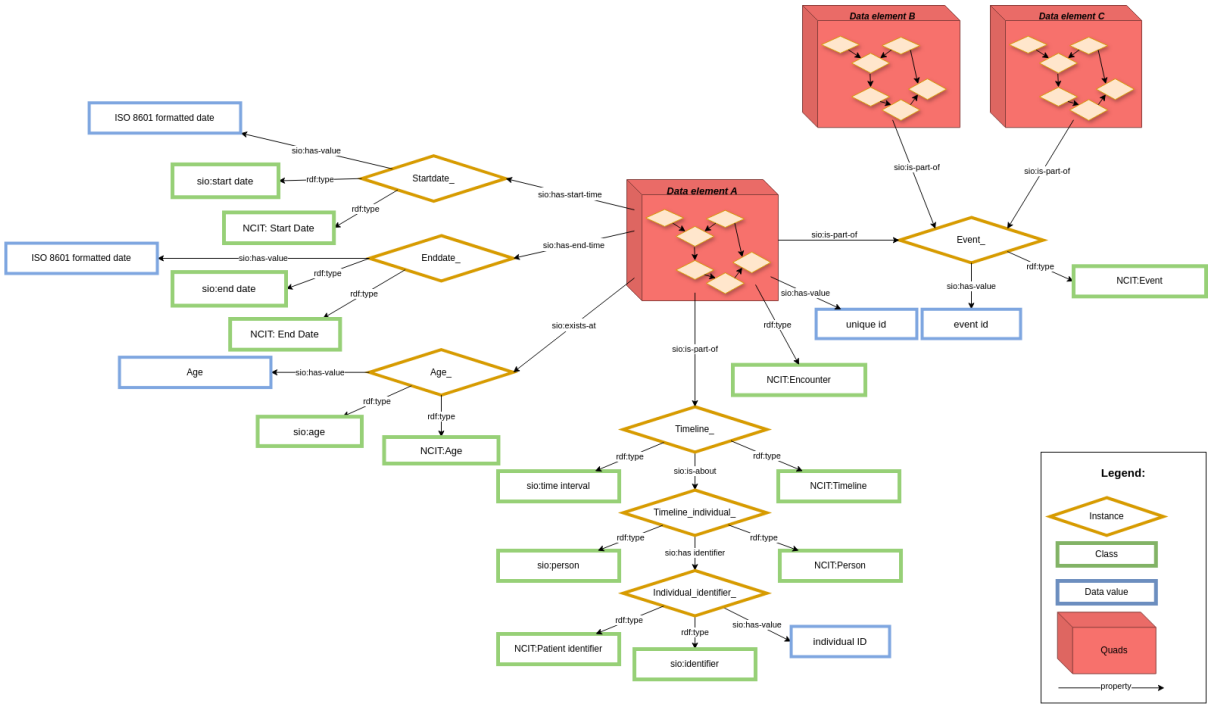


Figure 9: Overview of the addition of a contextual and temporal description layer to the model.

In order to solve these limitations, a metadata layer was created that imparts context on each data element. Semantically, the contextual metadata layer groups all components of an observation into a "named graph" which, itself, may have additional metadata facets. In RDF, context is modeled using RDF-Quads. This context URI can then be used as the subject for additional triples in order to, for example, add temporal or administrative information about that data element, or to group sets of triples into other higher-level structures.

An example of the use of named graphs is shown in Figure 9. The red boxes refer to a single data element that has been "encapsulated" in a common contextual named graph. In this example, the purpose of the named graph is to allow for the addition of temporal information in the form of time points or time intervals. As additional annotations on the named graph, this red box *is part of* a patient timeline description and an event encounter description. The use of an encounter identifier can be added to the model to further relate several of these data elements under the same clinical episode or event, for example, a treatment regimine.

Through the "context" node of RDF-Quads, arbitrary data elements can be linked, for example, the multiple data elements that arise from a single patient encounter with the healthcare system. Few resources seem to be taking advantage (in the rare disease space) of this RDF-Quad technology, despite it being a well-documented and official W3 standard for RDF representation for about a decade. Unfortunately, we had limited opportunity to study the benefits, since this addition came late in the project, and rich annotation of named graphs is not part of the FiaB ETL workflow.

5.3.6. Building a new core model

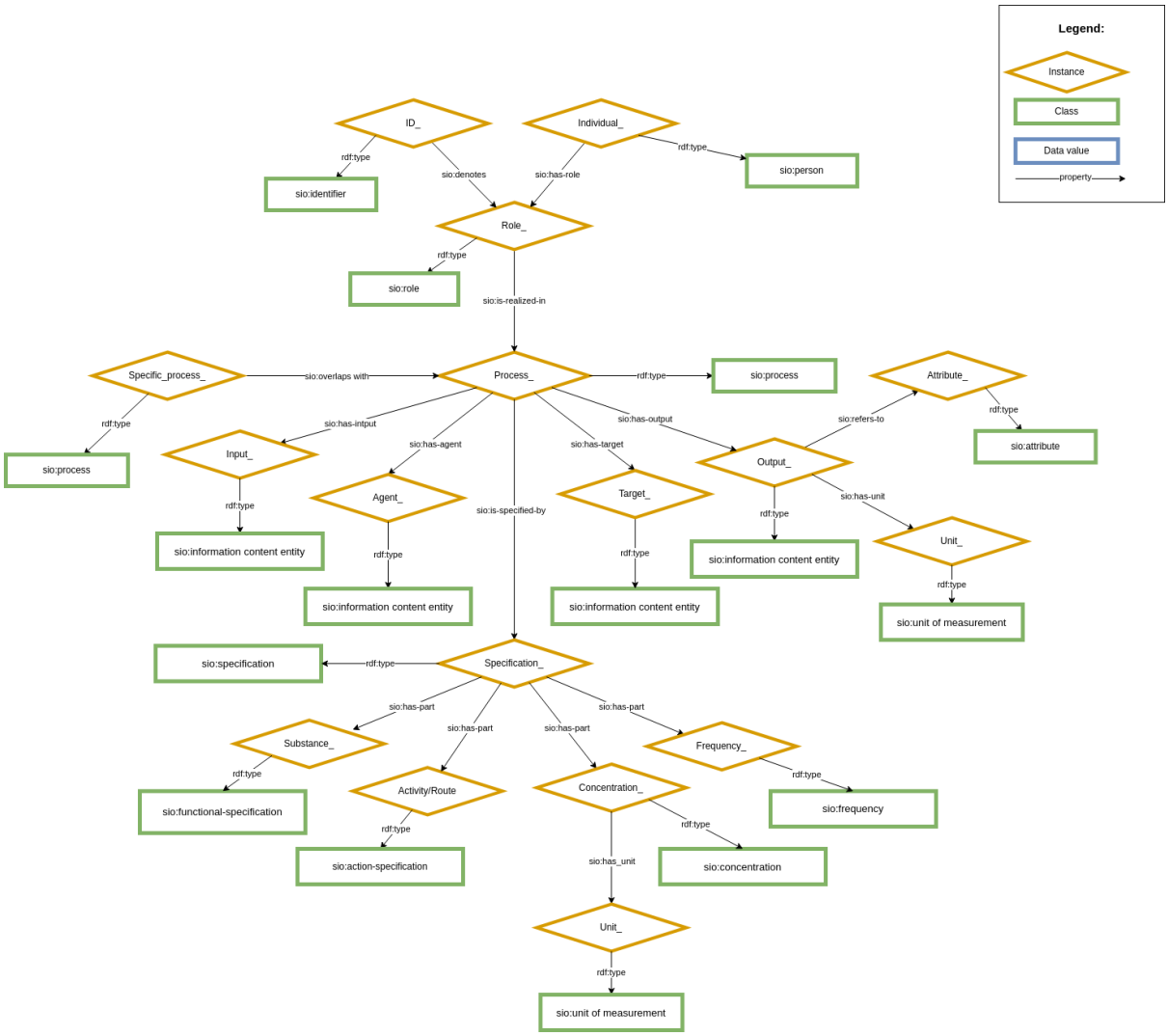


Figure 10: New core structure for the semantic data model.

In collaboration with Dr. Michel Dumontier (author of the SIO ontology) we undertook to rebuild the core data model for EJP-RD such that it could accurately represent all of the novel data requirements described above and address each limitation.

Of particular relevance were the following additions to the core model, as shown in Figure 10. The process - for instance, a clinical procedure - is now related to several additional entities beyond the process output, including inputs, agents, protocols and targets.

Input, agent and target are new sub-components added in the form of **information content entity** in the assessment process. *Has input*, *has agent* and *has target* are properties in SIO used for representing relationships between a **process** and this **information content entity**.

Protocol, typed as a "specification", is described by the use of other information content entities that are parts of this specification. These entities describe:

- **Action specification** such as activity or route of administration
- **Frequency** that the protocol defines, for instance, Medication drug frequency
- Chemical substance, defined as **function specification**
- **Concentration**, followed by its **unit of measurement**

User-feedback also suggested the requirement for additional types of processes. For some measurement representations, such as genetic testing or laboratory analyses, the specific type of test must be indicated in order to properly interpret the output. As with previous modelling attempts, the general process type (from SIO) is retained for the purpose of general queries, and a second more specific process type is added to the same node to support more specific queries.

Not all of the components added to the new model are required for every data element. Distinct data elements will use different combinations of these new components depending on the element being modeled. All the new individual data data element representations are in Annexes B.

5.3.7. EJP-RD Project Phase 2

These changes to the core model were then applied in response to feedback from the European Resource Networks (ERNs), who needed to describe a wider range of clinical measurements and observations. These new clinical element types included:

- Patient body measurements
- Laboratory molecular analyses
- Medical imaging techniques
- Drug medication prescriptions
- Surgical Medications
- Questionnaire information, formatted as Patient Reported Outcomes (PROs) (Weldring & Smith, 2013).

We will now describe these new models in some detail, where we use boldface to indicate **classes**, and italics to indicate *relations*.

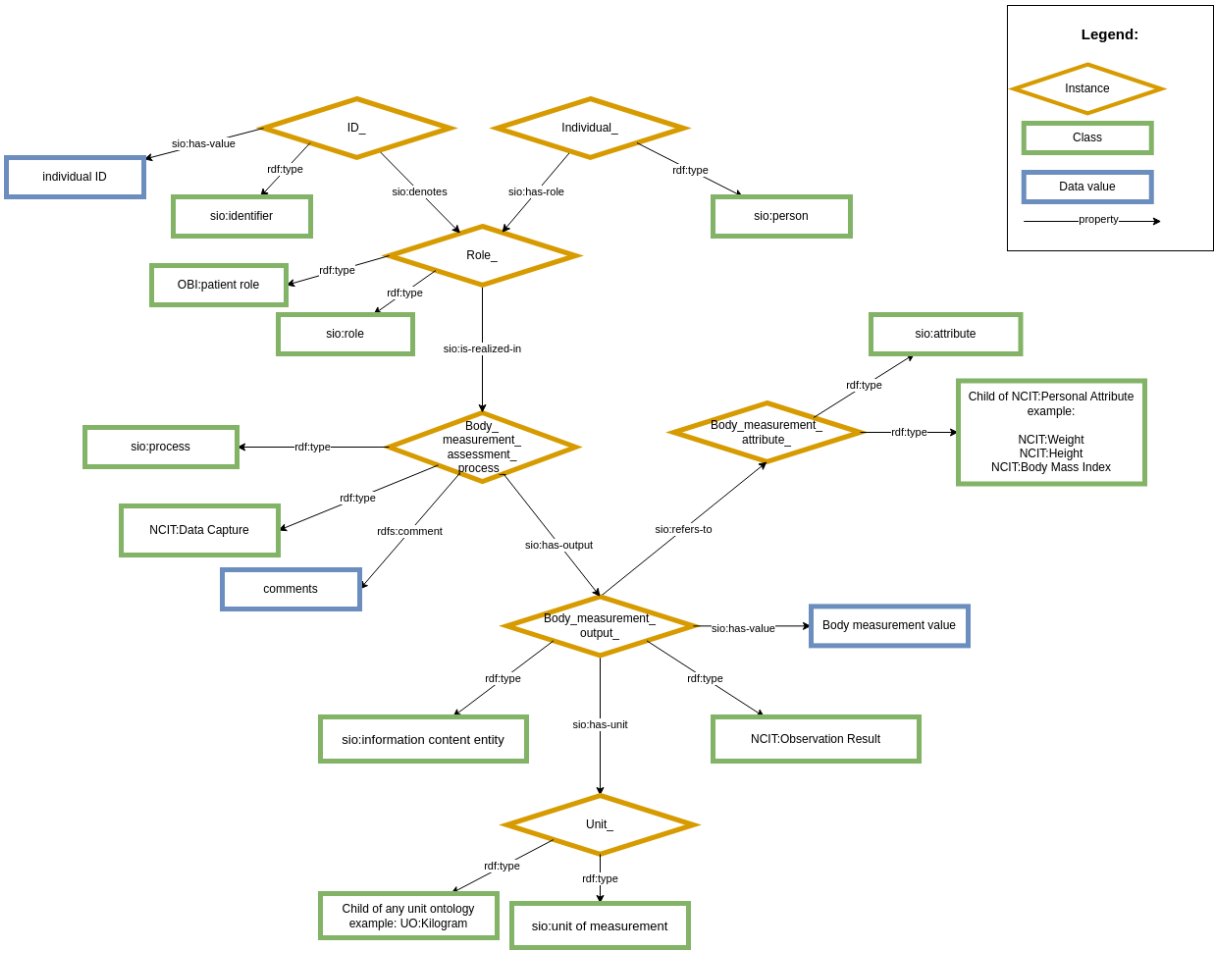


Figure 11: Body measurements data element.

Figure 11 describes a **process** of estimation of a corporal measurement. This **process** *has output* a measurement value, represented as an **information content entity**, along with its **unit of measurement**. This output refers to a patient **attribute**, which is the quality the corporal measurement is describing, in this particular case, **weight**. The definition of weight, or any other possible corporal attribute (for instance, height or Body Mass Index) is represented with the NCIT child term of **Patient Attribute.**

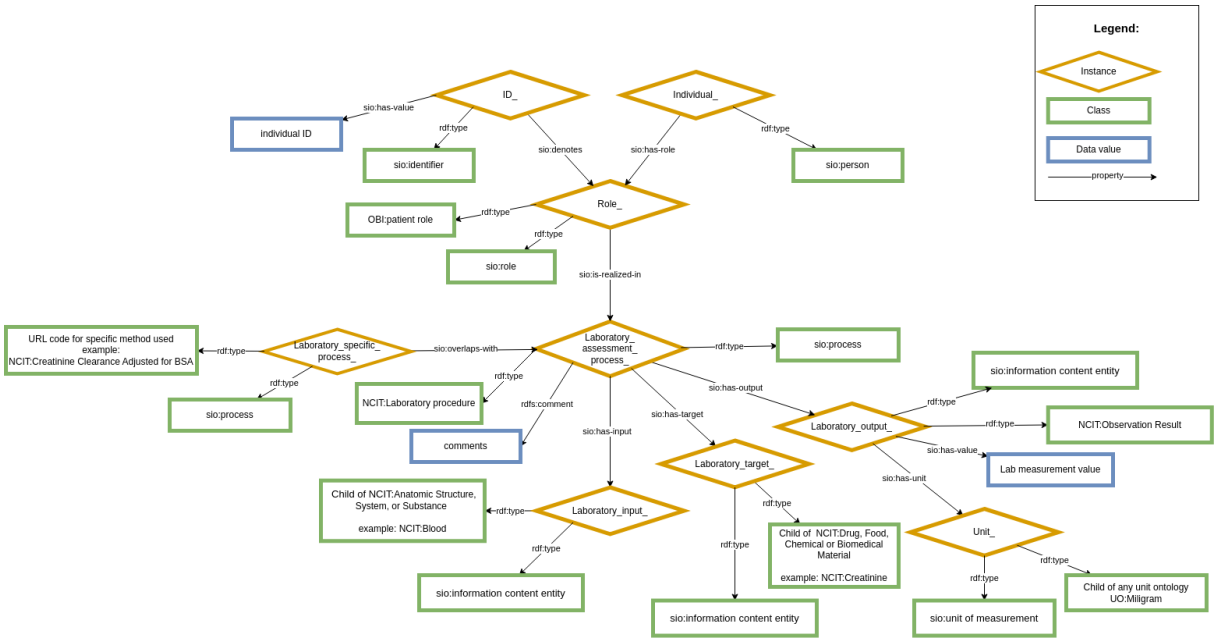


Figure 12: Laboratory molecular analysis data element.

Figure 12 describes a laboratory measurement **process**. In this **process,** it *has output* a measurement value, represented as an **information content entity**, along with its **unit of measurement.** Other participants are described at this assessment, it *has input* another **information content entity**, representing the anatomical structure where this sample is obtained. Moreover, it *has target* another **information content entity**, describing the molecular substance, which is a sample analyte.

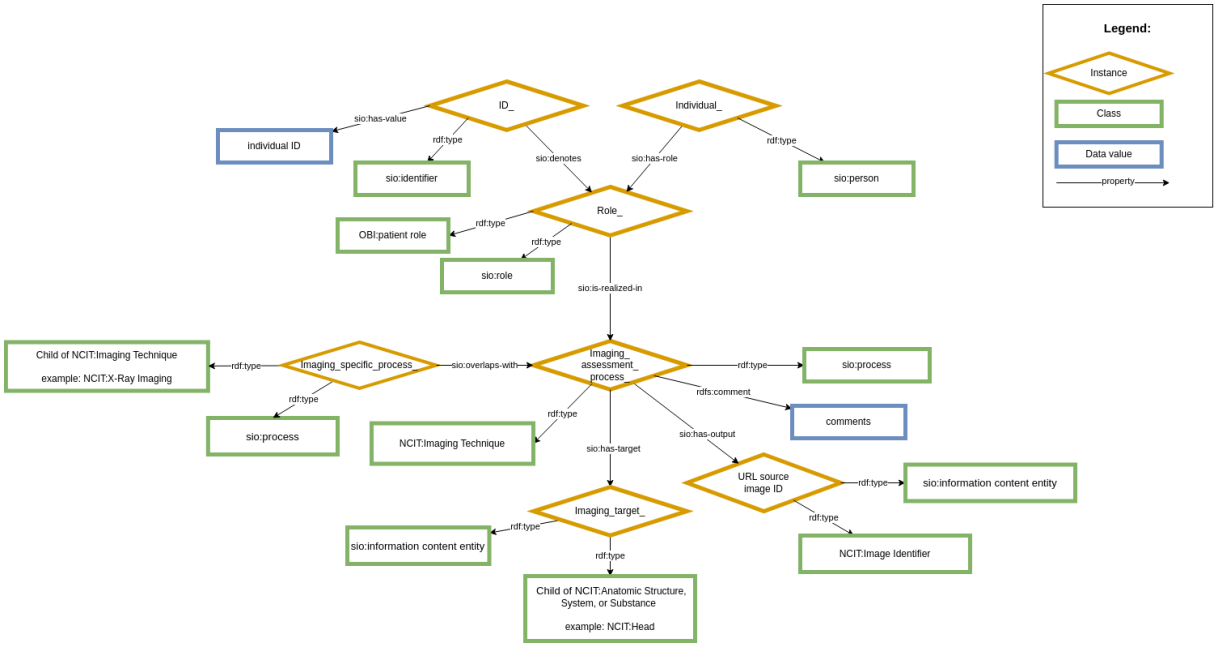


Figure 13: Medical imaging data element.

Figure 13 describes a medical technique **process**. In this **process,** it *has output* a measurement value, represented as an **information content entity**, describing the medical image identifier. Another participant is described at this assessment, it *has input* another **information content entity**, representing the anatomical structure where this medical image was taken.

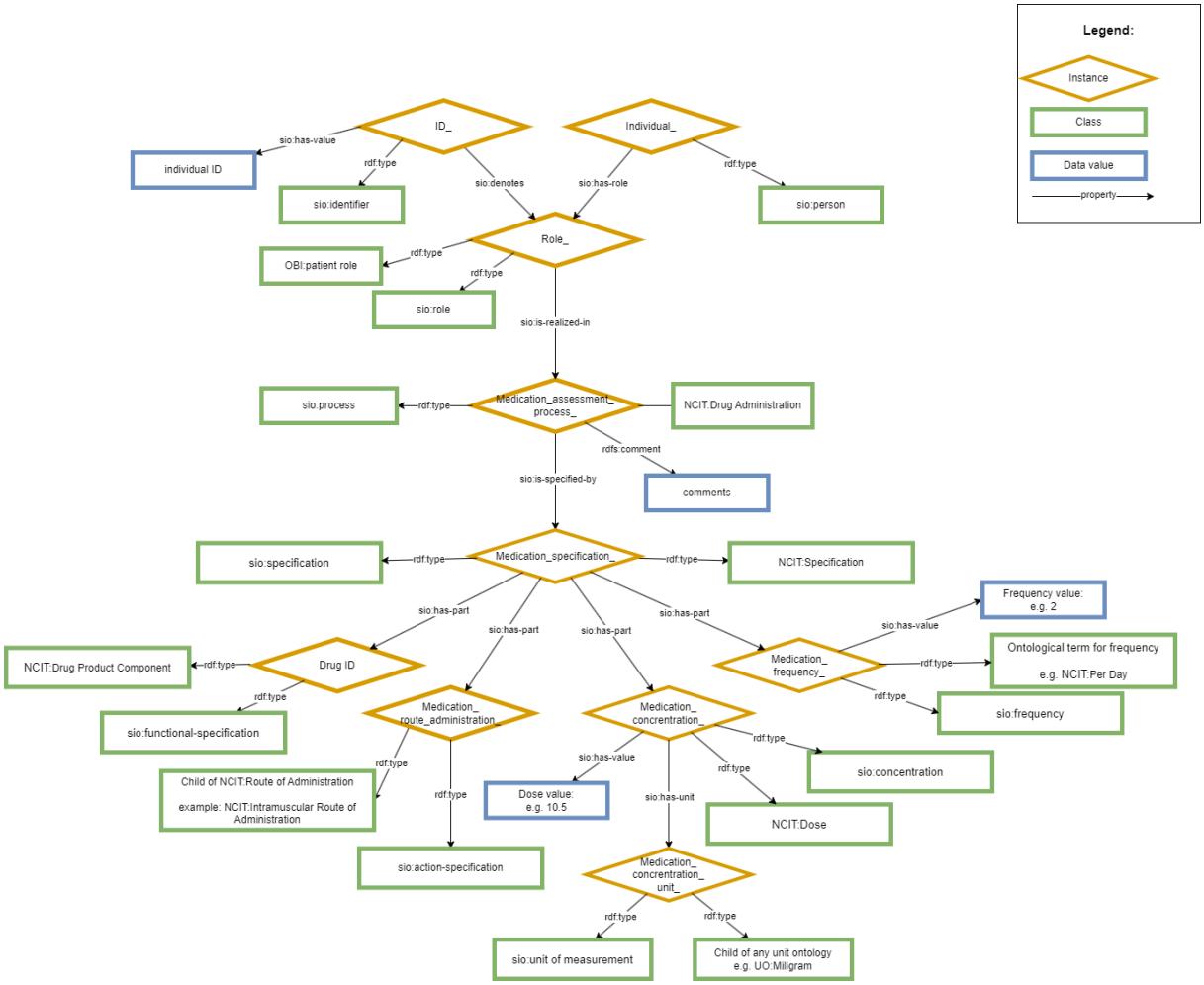


Figure 14: Medication data element.

Figure 14 describes a Drug administration procedure, represented by the **process** node. The process is *specified by* an administration protocol, defined as **specification** This **specification** contains several descriptions included, for instance: the consumed drug identifier, frequency and route of the drug administration, concentration, followed by its concentration units.

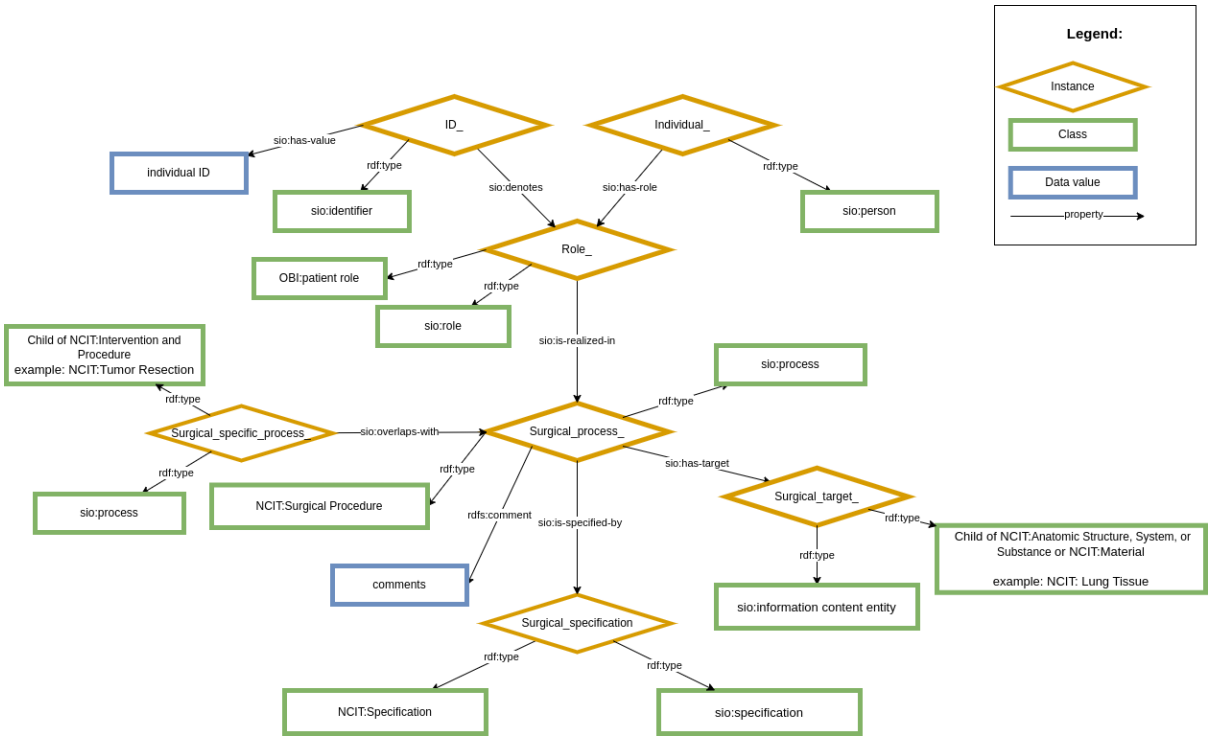


Figure 15: Intervention data element.

Figure 15 describes a Surgical procedure, along with its intervention procedure, for example, a Tumor resection, represented by the **process** node. The **process** is *specified by* an intervention **specification** (representing the protocol that was followed).

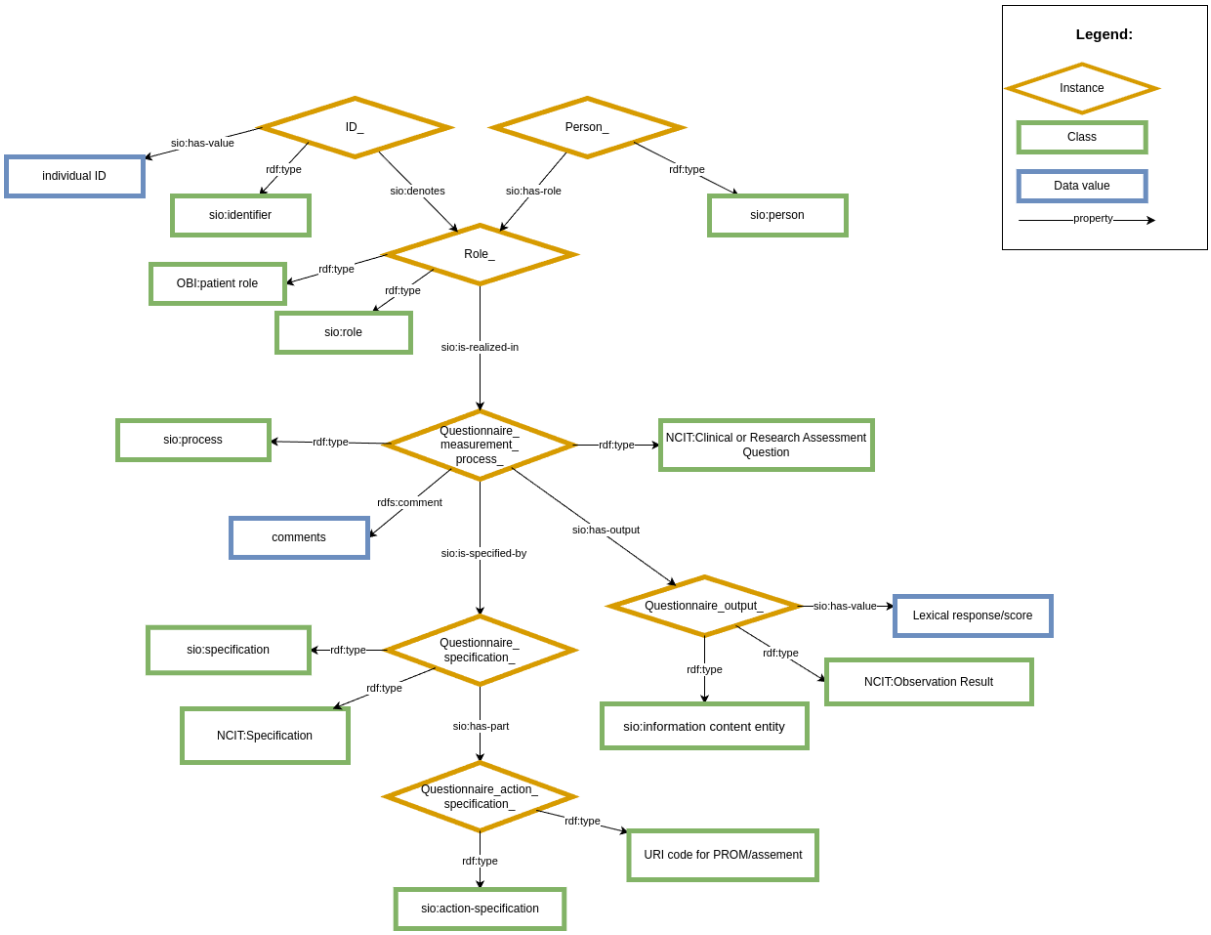


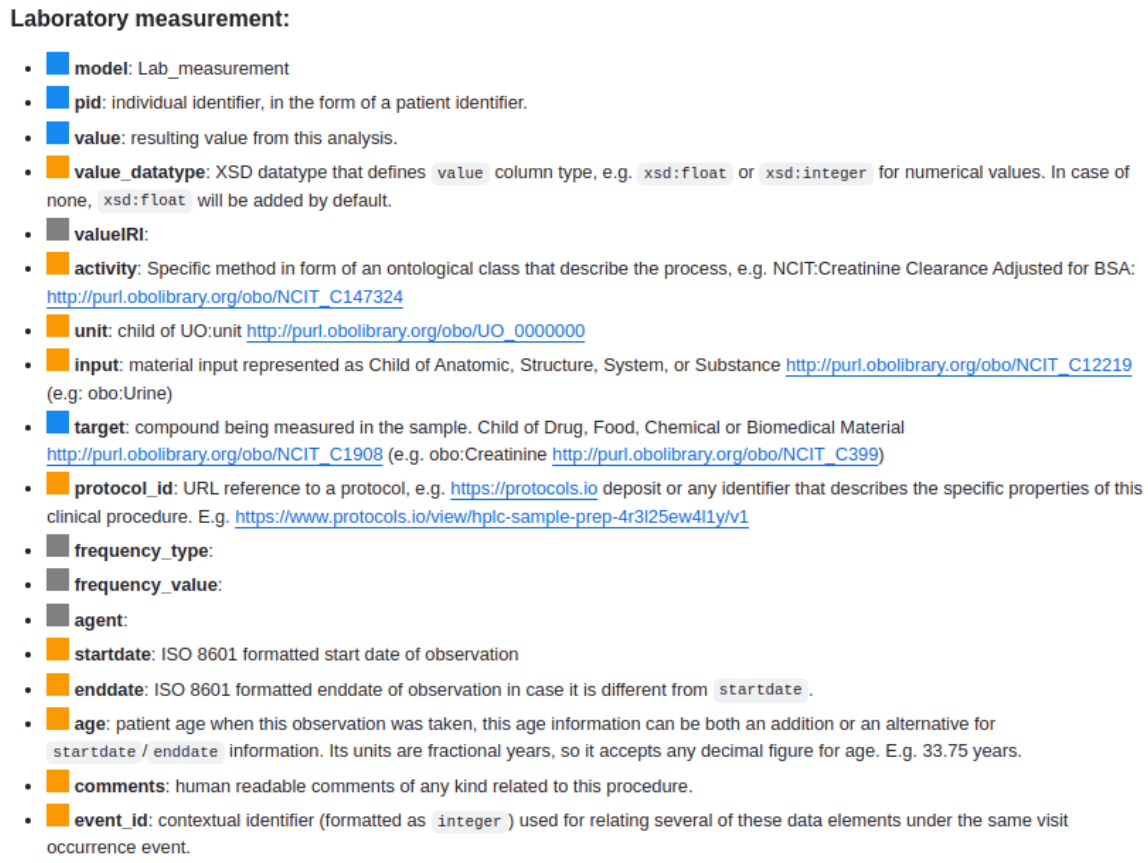
Figure 16: Questionnaire data element.

Figure 16 describes a process where a clinical research question is asked, represented by the **process** node. The process *is specified by* a questionnaire protocol, defined as **specification**. This protocol is, for example, a Patient Reported Outcome, defined as an **action specification** for this particular clinical question.

With the expansion of the scope of this core model, and the extension of the model into a wide range of novel clinical data types beyond the RD Platform CDEs, this model is no longer limited to common data elements. A new model name was chosen - **Clinical And Registry Entries Semantic Model[22] (CARE-SM)** - to reflect the new scope of healthcare data that can be represented by this model.

[22] https://github.com/CARE-SM/CARE-Semantic-Model

5.3.8. Creating a toolkit for CARE-SM implementation and data pre-evaluation



Laboratory measurement:

- **model**: Lab_measurement
- **pid**: individual identifier, in the form of a patient identifier.
- **value**: resulting value from this analysis.
- **value_datatype**: XSD datatype that defines `value` column type, e.g. `xsd:float` or `xsd:integer` for numerical values. In case of none, `xsd:float` will be added by default.
- **valueIRI**:
- **activity**: Specific method in form of an ontological class that describe the process, e.g. NCIT:Creatinine Clearance Adjusted for BSA: http://purl.obolibrary.org/obo/NCIT_C147324
- **unit**: child of UO:unit http://purl.obolibrary.org/obo/UO_0000000
- **input**: material input represented as Child of Anatomic, Structure, System, or Substance http://purl.obolibrary.org/obo/NCIT_C12219 (e.g: obo:Urine)
- **target**: compound being measured in the sample. Child of Drug, Food, Chemical or Biomedical Material http://purl.obolibrary.org/obo/NCIT_C1908 (e.g. obo:Creatinine http://purl.obolibrary.org/obo/NCIT_C399)
- **protocol_id**: URL reference to a protocol, e.g. https://protocols.io deposit or any identifier that describes the specific properties of this clinical procedure. E.g. https://www.protocols.io/view/hplc-sample-prep-4r3l25ew4l1y/v1
- **frequency_type**:
- **frequency_value**:
- **agent**:
- **startdate**: ISO 8601 formatted start date of observation
- **enddate**: ISO 8601 formatted enddate of observation in case it is different from `startdate`.
- **age**: patient age when this observation was taken, this age information can be both an addition or an alternative for `startdate` / `enddate` information. Its units are fractional years, so it accepts any decimal figure for age. E.g. 33.75 years.
- **comments**: human readable comments of any kind related to this procedure.
- **event_id**: contextual identifier (formatted as `integer`) used for relating several of these data elements under the same visit occurrence event.

Figure 17: Exemplar CSV template documentation for defining which columns to provide. This documentation corresponds to the Laboratory measurement description. As a legend. Blue boxes describe columns that are mandatory, orange boxes those that are optional and grey boxes, unused columns for this representation.

The CARE-SM model is more internally consistent, but its complexity has been increased. This required us to adapt the model implementation for both YARRRML and CSV templates to keep the execution of the ETL workflow sufficiently straightforward for our user community. Changes in the YARRRML have been described in the previous section, now we will focus on describing the adaptations of the CSV data source files. The following adaptations are managed by a new component of the ETL pipeline called the CARE-SM Toolkit[23], which has as input a minimal CSV file created by the user, and as output a much richer CSV containing additional columns required to properly fill the models. The Toolkit functionalities include:

[23] https://github.com/CARE-SM/CARE-SM-Toolkit

1. The independent YARRRML templates used for CDE-SM transformation were adapted and consolidated. In the resulting unified YARRRML template, different datatypes are managed by the CARE-SM Toolkit, via manipulating the input CSV in a datatype-specific manner. Thus, unlike the CDE-SM transformations, the CSV created by the end-user is not the raw input to the ETL pipeline, rather it is the modified CSV that has been automatically enhanced by the Toolkit.

2. The CARE-SM Toolkit populates the CSV template with all the OBO classes that are not included in the YARRRML template, thus the user does not need to select appropriate ontological terms when the term has been pre-defined. The addition of the appropriate ontological concepts is controlled by a "tagname" - in one of the CSV columns - that specifies which data element that row is intended to represent. The Toolkit is thus able to add the appropriate terminology in its output CSV.

3. Although the core structure is unified and all data elements are modeled individually, data derived from the source registry is positioned in different sub-components depending on which model is being instantiated. For instance, height is a measurement value, whereas diagnosis is defined as a condition (Personal Attribute). The CARE-SM Toolkit recognizes (via the tagname) the data element modeled and locates the data information into the proper sub-component reference in the template. Thus, the same "value" column in the template can be reused for multiple models, simplifying the template.

4. Quality control over the completed CSV, checking for unacceptable columns that arise through mis-interpretation of the documentation or typographical errors. Additionally, quality control is conducted on time interval-related columns, such as startdate and enddate ensuring they adhere to the ISO-8601 notation standard. Also, if the enddate is not provided by the user, the information in the startdate column is copied by the Toolkit into the enddate column to complete the time interval, simulating a time point.

5. The elimination of any data representation that lacks the minimum requirement for its serialization, avoiding the generation of incomplete RDF. A glossary of all data elements is documented in Github, describing mandatory and optional CSV columns to populate for every case, shown in Figure 17. CARE-SM Toolkit will erase every row that doesn't pass the specific requirements of each data element. A final report to the user detailing the removed rows is generated after this data quality control step.

6. The inclusion of a script for generating unique timestamps. These timestamps are used for creating unique identifiers (URLs) for every instance in the model, ensuring that the RDF instances are guaranteed to be distinct.

5.4. Can we perform federated data exploration using our semantic data model?

5.4.1. Implementation and automated deployment of the Beacon-2 API for data discovery

By internal agreement in EJP-RD, the Beacon version 2 specification (Rambla et al., 2022) has been adopted for resource discovery with regard to data catalogs, registry datasets, and biosamples. Beacon is a REST API defining the content of HTTP-based request/responses that fosters homogenization of data exploration over diverse types of clinical databases. It uses a JSON syntax to define all request parameters including a filter that enables the message to define query constraints. The response, also formatted in JSON, is the count of the number of entities that match against the filtered query. In this way, it is privacy-preserving as it does not expose any personal information, rather it provides only aggregate counts.

CARE-SM concept	Ontological Term	Beacon-2 Filter Type	ID	Operator	Permitted values
Sex	ncit:C28421	Alphanumeric	ncit:C28421	=	Any array of the following terms: ncit:C16576 ncit:C20197 ncit:C124294 ncit:C17998
Disease or Disorder	ncit:C2991	Ontology	A single value or an array of orphanet terms. e.g. ordo:Orphanet_558 or [ordo:Orphanet_558, ordo:Orphanet_773]	NA	NA
Phenotype	sio:SIO_010056	Ontology	A single value or an array of HPO terms. e.g. hp:0001251 or [hp:0001251, hp:0012250]	NA	NA
Causative Gene	edam:data_2295	Alphanumeric	edam:data_2295	=	any HGNC gene symbol or array of HGNC symbols
Year at birth	ncit:C83164	Numerical	ncit:C83164	=, >=, >, <=, <	Any integer

Symptom Onset	ncit:C124353	Numerical	ncit:C124353	=, >=, >, <=, <	Any integer
Age at diagnosis	ncit:C156420	Numerical	ncit:C156420	=, >=, >, <=, <	Any integer

Table 6: Alignment of the CARE-SM data elements, their associated Beacon-2 filters, and their potential filter values.

The Beacon team, in collaboration with EJP-RD, defined a set of filters for privacy-preserving queries over patient information. In combination with CARE-SM, this set of filters can be used for discovery of resources that contain records consistent with the filters. These filters take three facets: value, operator, and the list of permitted values, shown in Table 6.

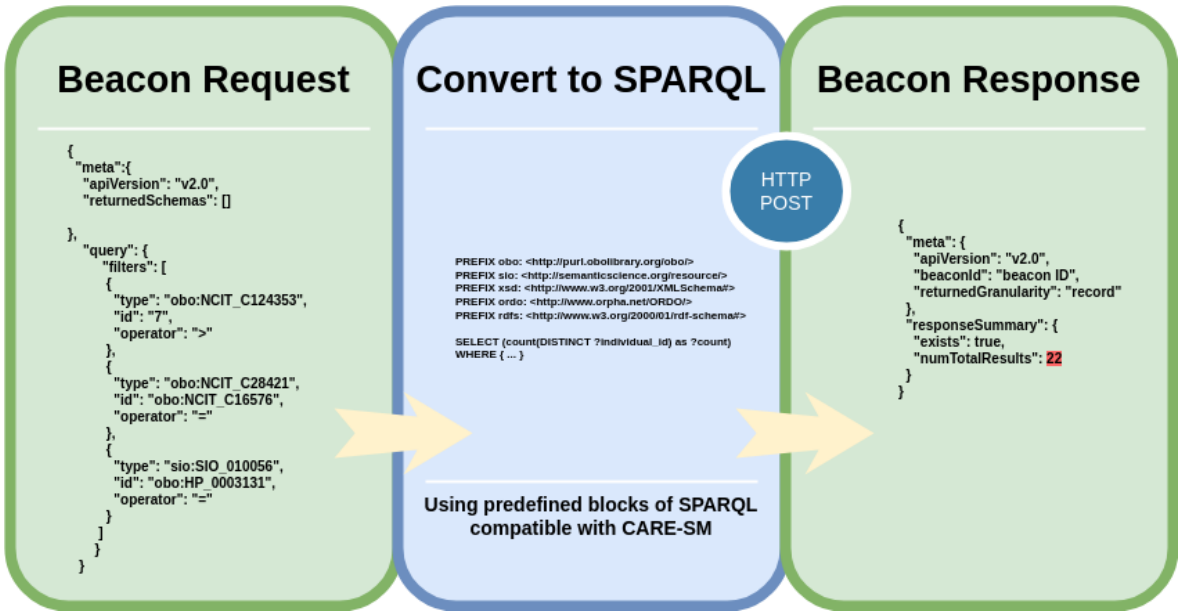


Figure 18: Beacon-2 API for CARE-SM workflow.

Because CARE-SM is a uniform data model, it was possible to automate the implementation of Beacon for querying CARE-SM formatted data; thus CARE-SM adopters do not need to create their own implementation of the Beacon-2 API, regardless of the structure of their native data store. To fulfill this objective, we convert the filters in the incoming Beacon JSON request into constraint clauses in a SPARQL query, and because of the predictability of the CARE-SM models these queries can be templated. The set of constraint blocks are joined into a final SPARQL "count" query which is passed to the Triplestore. The resulting count number is placed into the Beacon JSON response, as shown in Figure 18 in the red box. No sensitive patient information is exposed. The whole implementation is documented in Guthub[24].

Apart from data query, other portions of the Beacon-2 API were implemented, such as the call to retrieve the available filters supported by the server. Finally, we implemented a metadata description of the Beacon-2 API using the OpenAPI metadata standard for describing Web services. The API was Dockerized and added to the FiaB installer for easy deployment.

5.4.2. Testing our Beacon-2 API implementation with CARE-SM patient data

The Beacon-2 API for CARE-SM was tested by the use of 100 patients from the synthetic ERKreg dataset, transformed into CARE-SM data elements. A set of these data elements aligned with the Beacon-2 list of filters from Table 6 was selected. This selection includes:

- Patient birthdate information
- Patient sex
- Methylmalonic acidemia with homocystinuria diagnosis, annotated by Orpha code: 79282
- Cystinuria phenotype, annotated by Human Phenotype Ontology: 0003131
- Age of the Methylmalonic acidemia with homocystinuria diagnosis onset
- Age of Cystinuria symptom onset

[24] https://github.com/CARE-SM/beaconAPI4CARESM

- Gene variant associated with Gene SLC3A1 for Cystinuria, code using ClinVar. HGVS: NM_000341.3(SLC3A1):c.-82T>G

This above patient information was serialized into RDF and stored in a triplestore, using it as the SPARQL endpoint. Beacon-2 API calls were sent to test the fidelity of the response over this "gold standard" dataset, where the correct result of every query was known.

Data element filtered	**Beacon-2 API JSON filter**

Female sex	{
	"type": "obo:NCIT_C28421",
	"id": "obo:NCIT_C16576",
	"operator": "="
	}

Methylmalonic acidemia with homocystinuria diagnosis after 15 years old	{
	"type": "obo:NCIT_C2991",
	"id": "ordo:Orphanet_79282",
	"operator": "="
	}, {
	"type": "obo:NCIT_C156420",
	"id": "15",
	"operator": ">"
	}

Cystinuria detected after 9 years old	{
	"type": "sio:SIO_010056",
	"id": "obo:HP_0003131",
	"operator": "="
	},{
	"type": "obo:NCIT_C124353",
	"id": "9",
	"operator": ">"
	}

Genetic examination of NM_000341.3(SLC3A1): c.-82T>G variant	{
	"type": "edam:data_2295",
	"id":"https://www.ncbi.nlm.nih.gov/clinvar/variation/897001/",
	"operator": "="
	}

Table 7: Beacon-2 API filter tested using mock patient data.

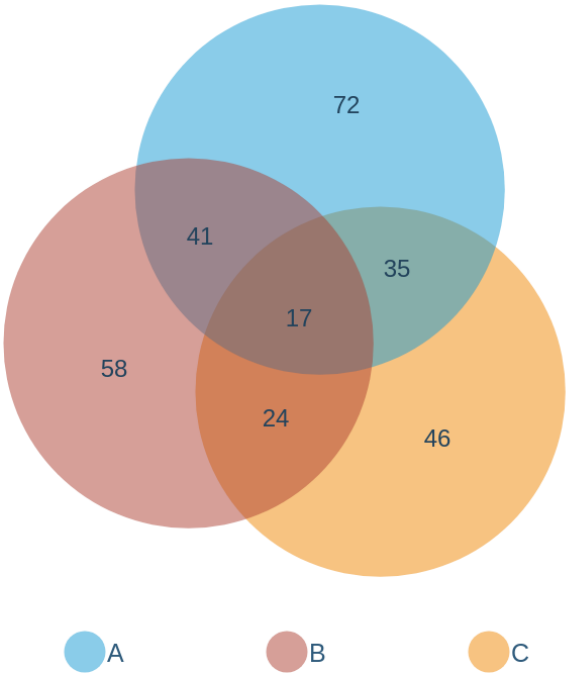


Figure: 19 Venn diagram for the resulting outcome of the combination of multiple Beacon-2 filters. (A) refers to being diagnosed of Methylmalonic acidemia with homocystinuria after 15 years old, (B) refers to Female sex and (C) refers to Cystinuria phenotype.

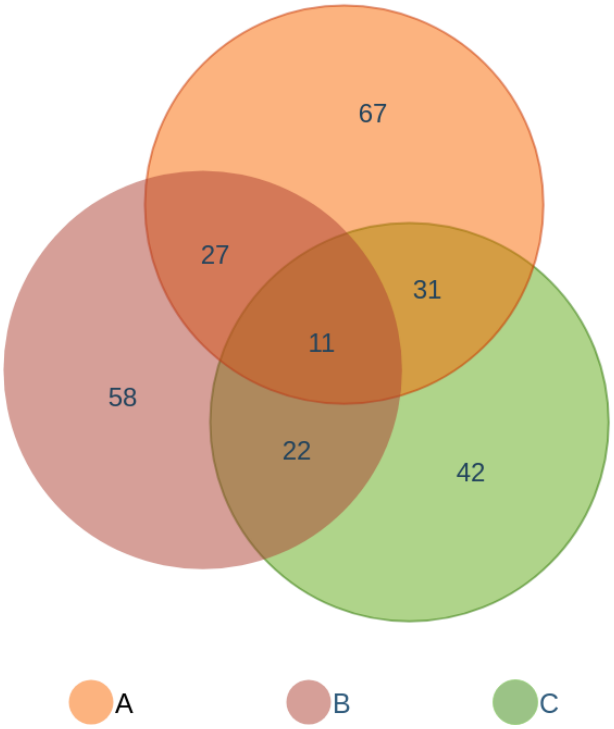


Figure: 20 Venn diagram for the resulting outcome of the combination of multiple Beacon-2 filters. (A) refers to NM_000341.3(SLC3A1):c.-82T>G variant , (B) refers to Female sex and (C) refers to symptom onset on Cystinuria after 7 years old.

All the filters from Table 7 were tested both individually and as permutations with other filters to test the fidelity of the automatically-generated SPARQL queries. In Figure 19 and 20, Venn diagrams describe all the combinations of filters performed against the mock patient data registry. Each of the combinations successfully retrieved the expected count.

5.5. Do the CDE-SM/CARE-SM models facilitate interoperability with other common clinical data frameworks?

5.5.1. Interoperability between CDE-SM and the C-Path Institute data model.

A collaborative experiment in data interoperability was defined between the EJP-RD project and Critical Path (C-Path) Institute located in Tucson, Texas. This experiment explored the value gained by FAIRification, and the necessity for schema harmonization between different data models given their mutual commitment to FAIR representations. Both EJP-RD and C-Path organizations provided complementary patient datasets with the purpose of exploring data interoperability between their respective data models, both using Linked data and Semantic Web technologies. C-Path created their own semantic data model, while our experiments were executing using the CDE-SM models. This experiment involved three primary activities:

- Selection of datatypes to be modeled, and discussion about the "nature" of those selected data types, *v.v.* what needed to be modelled for the experiment.
- Identifying Biolink classes that match the concepts in both the EJP and C-PATH data models, such that each class could be "anchored" into a common conceptual layer.
- Querying both data models under a common SPARQL query that uses Biolink classes across both data models.

5.5.1.1. EJP-RD and C-Path datasets

The C-Path dataset contains aggregated data sourced from multiple studies gathered by the Polycystic Kidney Disease (PKD) Outcomes Consortium. C-Path data was used for developing CDISC data standards for PKD and have supported the qualifications of Total Kidney Volume as an imaging biomarker for drug development tools by regulatory agencies such as the Food and Drug

Administration (FDA) and the European Medicines Agency (EMA). The dataset, already anonymized, was further protected for this study through the synthesis of values for laboratory test results using the "synthpop" R package. From the EJP-RD side, this experiment reused the synthetic ERKreg-based dataset described previously.

Domain	C-Path	EJP-RD
Birthdate/Age patient information.	Age as an integer	ISO 8601 compliant date string
Sex patient information.	Sex label as a string (F, M) mapped to NCIT terms for Female and Male)	NCIT (National Cancer Institute Thesaurus) term for Sex (Female, Male, Undetermined or Unknown) and Sex label as a string.
Laboratory data measurements.	• Laboratory test name (e.g., Leukocytes), category (e.g., hematology), and specimen type (e.g., blood). • Numerical result and standard ranges in both original and standard units. • Study day of lab test. • Associated subject visit. • Associated specimen collection procedure.	• Procedure defined as Quantitation or Estimation. • Materials tested input • Target molecular or compound measured. • Output measurement value and its unit. • ISO 8601 compliant date of measurement procedure.

Table 8: Selected clinical information types present in both datasets.

After an exploration of the common variables defined in both datasets, 3 significant data domains related to PKD were selected as initial targets for testing federated data exploration, described in Table 8.

5.5.1.2. Identifying common Biolink classes for the data models

Biolink model entities	C-Path SDTM mapping	EJP-RD
Case https://w3id.org/biolink/vocab/Case	Subject https://w3id.org/c-path/biolink_sdtm_owl/SUBJECT	Person http://semanticscience.org/resource/SIO_000498
Procedure https://w3id.org/biolink/vocab/Procedure	Laboratory Test https://w3id.org/c-path/biolink_sdtm_owl/LBTEST and Urinary System Test https://w3id.org/c-path/biolink_sdtm_owl/URTEST	Process http://semanticscience.org/resource/SIO_000006
Information Content Entity https://w3id.org/biolink/vocab/InformationContentEntity	Information Content Entity https://w3id.org/biolink/vocab/InformationContentEntity	Information Content Entity http://semanticscience.org/resource/SIO_000015
Attribute https://w3id.org/biolink/vocab/Attribute	Does not exist in the PKD dataset. We reuse https://w3id.org/biolink/vocab/Attribute	Attribute http://semanticscience.org/resource/SIO_000614
Biological Sex https://w3id.org/biolink/vocab/BiologicalSex	Sex https://w3id.org/c-path/biolink_sdtm_owl/SEX	Sex http://purl.obolibrary.org/obo/NCIT_C28421

Table 9: Mapping of similar conceptual entities between Biolink, C-Path, and EJP-RD.

Table 9 shows the final mapping between Biolink concepts, and approximately matching concepts from each of the participating datasets, with the goal of using. Biolink as a "bridge vocabulary" between data models.

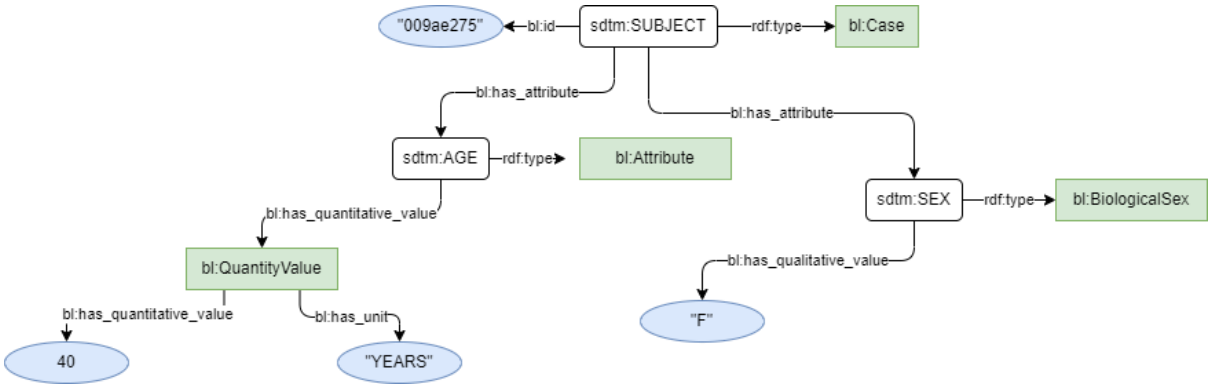


Figure 21: C-Path Patient information with the addition of Biolink (bl:) classifications.

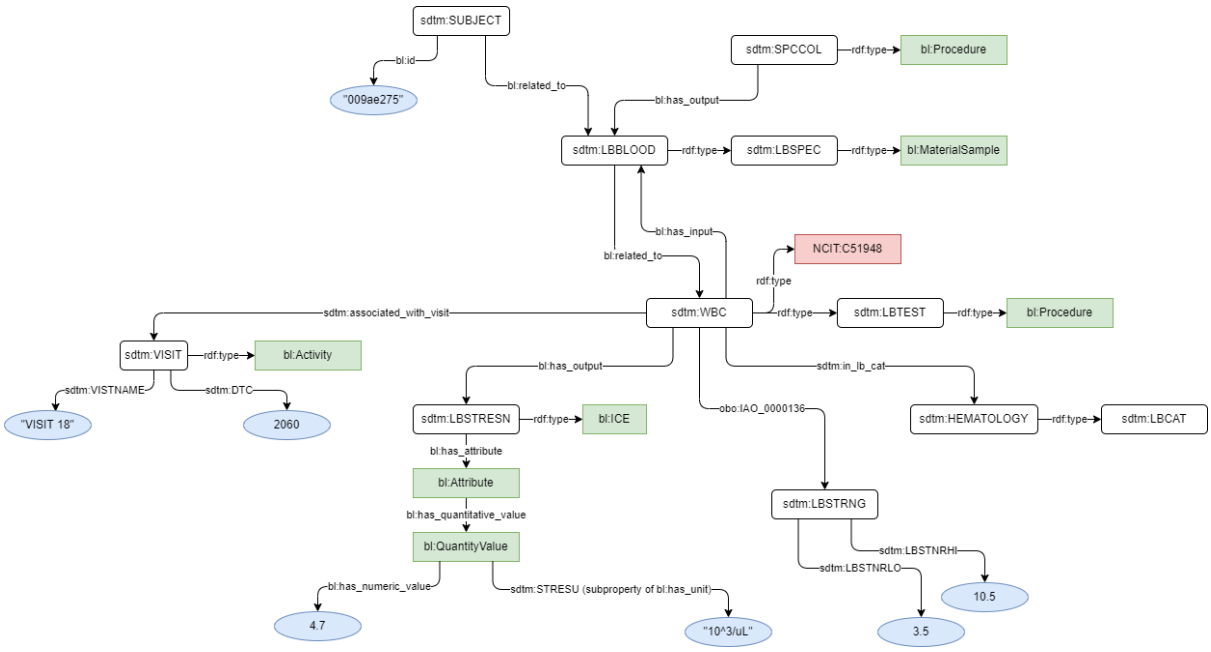


Figure 22: C-Path Leukocyte count measurement with the addition of Biolink (bl:) classifications.

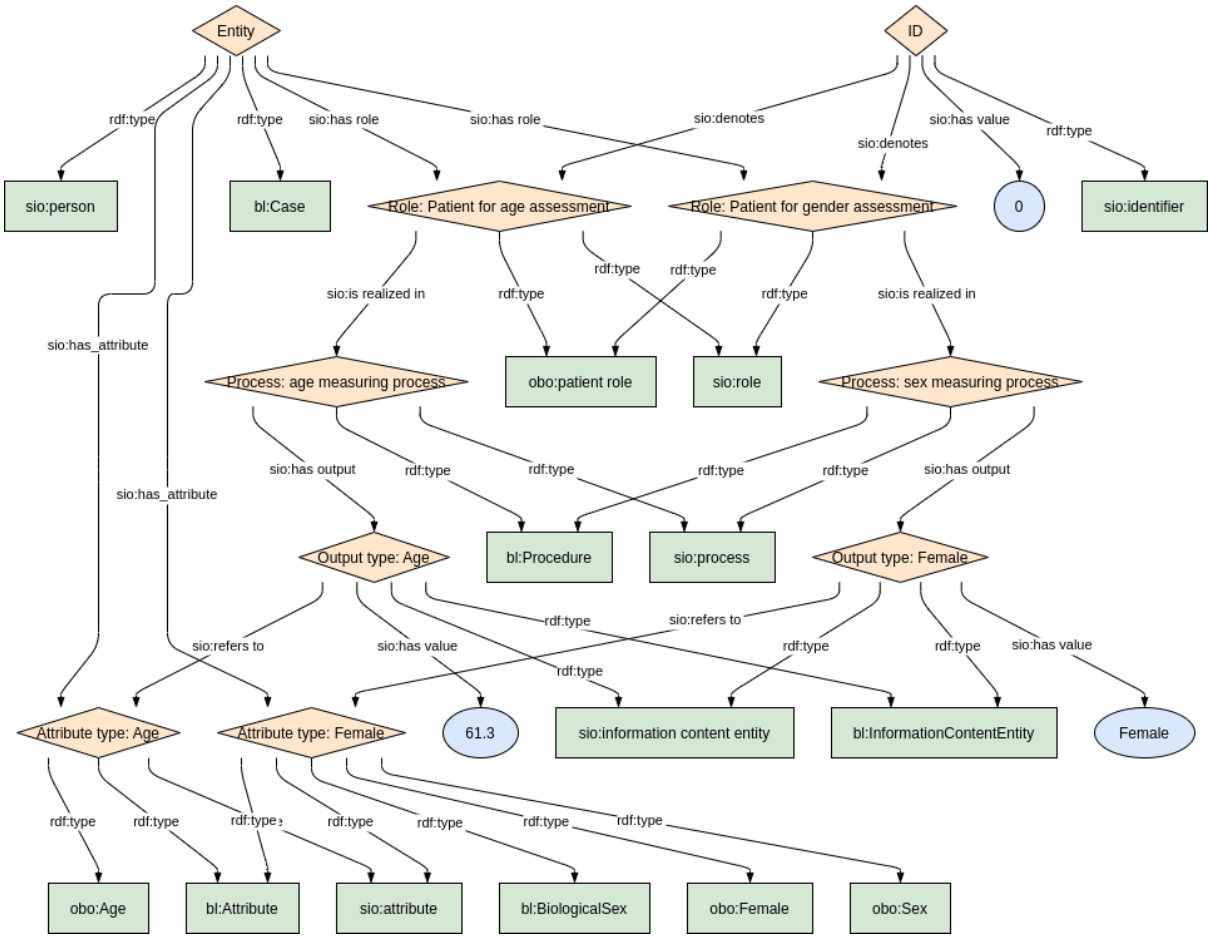


Figure 23: CDE-SM Patient information with the addition of Biolink (bl:) classifications.

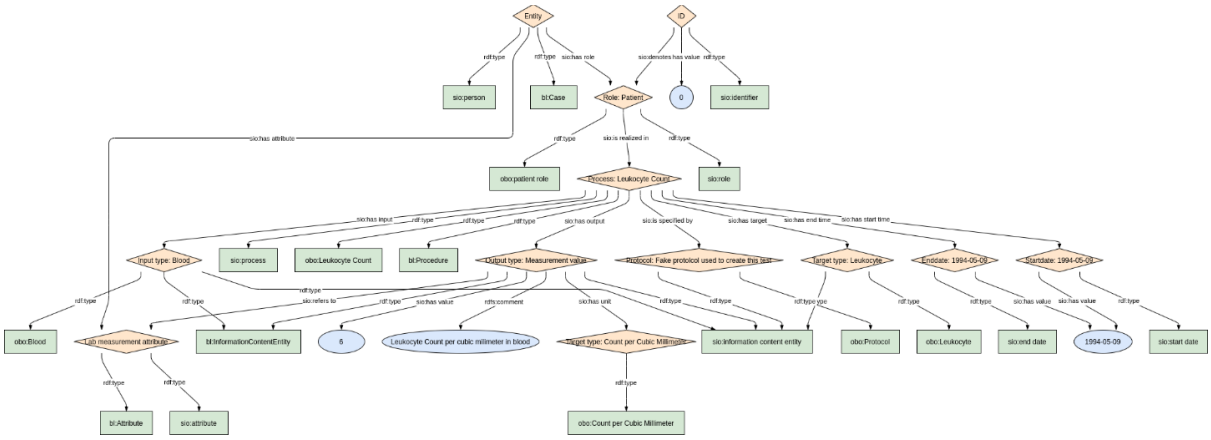


Figure 24: Leukocyte count measurement CDE-SM with the addition of Biolink (bl:) classifications.

The C-Path data model is designed by the combination of CDISC Study Data Tabulation Model (SDTM) (Wood & Guinter, 2008) and the Biolink Model. SDTM is a data model meant for collection, management, analysis and reporting clinical trial data. Figures 21 and 22 show how the C-Path data model describes the selected domains for this experiment, including the shared Biolink classes. Figures 23 and 24 show the representation of CDE-SM - including equivalent Biolink classes - of the same domain.

Data from each participant were serialized to RDF using YARRRML templates representing the semantic models of each site. After RDF generation, linked data was hosted in independent triplestores.

5.5.1.3. Federated queries over both data models

Query 1

```
PREFIX ncit: <http://purl.obolibrary.org/obo/>
PREFIX biol: <http://purl.org/NET/biol/ns#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX sio: <http://semanticscience.org/resource/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX biolink: <https://w3id.org/biolink/vocab/>
PREFIX bl: <https://w3id.org/biolink/>
PREFIX blowl: <https://w3id.org/c-path/biolink_sdtm_owl/>

SELECT DISTINCT ?test ?value WHERE {
 GRAPH <http://w3id.org/FAIR_Training_LDP/DAV/home/LDP/cpath/cpath_full> {
 ?test a biolink:Procedure, ncit:NCIT_C51948 .
 ?test ?has_output ?output .
 ?output a biolink:InformationContentEntity .
 ?output bl:has_attribute ?att .
 ?att bl:has_quantitative_value|bl:has_qualitative_value ?valnode .
 ?valnode bl:has_numeric_value ?value
 }
}
```

514 Results

Query 2

```
PREFIX ncit: <http://purl.obolibrary.org/obo/>
PREFIX biol: <http://purl.org/NET/biol/ns#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX sio: <http://semanticscience.org/resource/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX biolink: <https://w3id.org/biolink/vocab/>
PREFIX bl: <https://w3id.org/biolink/>
PREFIX blowl: <https://w3id.org/c-path/biolink_sdtm_owl/>

SELECT ?value ?unit WHERE {
 GRAPH <http://w3id.org/FAIR_Training_LDP/DAV/home/LDP/cpath/cbgp_leuk> {
 ?test a biolink:Procedure, ncit:NCIT_C51948 .
 ?test ?has_output ?output .
 ?output a biolink:InformationContentEntity .
 ?output sio:SIO_000300 ?value .
 ?output sio:SIO_000221 ?unitnode .
 ?unitnode rdfs:label ?unit
 }
}
```

3554 Results

Query 3

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX biolink: <https://w3id.org/biolink/vocab/>
PREFIX bl: <https://w3id.org/biolink/>

SELECT DISTINCT ?test ?value ?unit WHERE {
{SERVICE <http://fairdata.systems:8890/sparql>{
 {SELECT ?test ?value where {
 GRAPH <http://w3id.org/FAIR_Training_LDP/DAV/home/LDP/cpath/cpath_full> {
 ?test a biolink:Procedure, ncit:NCIT_C51948 .
 ?test ?has_output ?output .
 ?output a biolink:InformationContentEntity .
?output bl:has_attribute ?att .
?att bl:has_quantitative_value|bl:has_qualitative_value ?valnode .
?valnode bl:has_numeric_value ?value
 }
 }
 }
 }
}
UNION
{SERVICE <http://fairdata.systems:8890/sparql>{
 {SELECT ?test ?value ?unit where {
 GRAPH <http://w3id.org/FAIR_Training_LDP/DAV/home/LDP/cpath/cbgp_leuk> {
 ?test a biolink:Procedure, ncit:NCIT_C51948 .
 ?test ?has_output ?output .
 ?output a biolink:InformationContentEntity .
?output sio:SIO_000300 ?value .
?output sio:SIO_000221 ?unitnode .
 ?unitnode rdfs:label ?unit
 }
 }
 }
 }
 }
}

4068 Results
```

Table 10: SPARQL queries for the 3 different experiments. Query 1: Leukocyte Counts from C-Path dataset. Query 2: Leukocyte counts from EJP-RD. Query 3 Leukocyte counts from both datasets.

To test interoperability between both data models, SPARQL queries - shown in Table 10- were defined. While the SPARQL query contains SERVICE clauses to simulate two distinct endpoints for federated query, the same objective is achieved (in this case) by holding the C-Path and CDE-SM data in separate named graphs, and then executing the federation by joins over those two graphs. By querying both repositories using the common Biolink classes as a "bridge" between the two models, it became possible to construct query clauses that extracted the data from each. All expected data records of leukocyte count analysis are retrieved by the Query 3. While successfully demonstrating federated query, it should be noted that the queries over the two models had to be individually constructed to align with each data model, and there was no clear way to automate this task. This is a significant limitation that will be addressed further in the Discussion section. This work is described in more detail in the following publication (Alarcon et al., 2023).

5.5.2. Interoperability between CARE-SM and OMOP-CDM

5.5.2.1. Propose and planification of this experiment

EJP-RD, in its attempt to ensure data sharing and interoperability with different standards, have dedicated an effort to create multiple interoperability across different data model. In this experiment we are going to describe one of these activities in the form of schema mapping between CARE-SM and one the healthcare data model standard, OMOP-CDM version 5. 3..

A workgroup was organized by the EJP-RD project, to explore interoperability between OMOP CDM and CARE-SM. Bi-weekly meetings were held to identify a roadmap through the mapping activities. These meetings were coordinated by a group of experts in several fields such as schema interoperability and data integration, clinical data and metadata modelers and FAIR data stewardship. This workgroup was focused on the following objectives:

1. The creation of a mapping table that contains the concept lists of each data model, and the closest match between them.

2. The creation of an Extract, Transform, Load (ETL) workflow capable of transforming data represented using CARE-SM into OMOP-CDM.

Hackathon activities were scheduled on a yearly basis where the group gathered to plan, implement, and evaluate solutions to this mapping exercise and to then define and codify a data transformation workflow.

5.5.2.2. Concept representation in both data models

Although both OMOP-CDM and CARE-SM implement Semantic Web technologies in their respective data models, both their schema and ontologies differ, resulting in a situation similar to the one we faced with CDE-SM and C-Path. Thus, we followed a similar approach in this study. A comparison of the choice of ontological annotations for similar concepts is required for devising first if there is any "*lingua franca*" between schemas, and if not, we must then define the similarity of the chosen concepts.

Data element	Ontological terms in CARE-SM	Ontological terms in OMOP-CDM	OMOP data table	OMOP column	Label	Terminology	Mapping type
Sex	NCIT:sex	Not Applicable	Person			SNOMED	
Sex	NCIT:female	8532	Person	gender_concept_id, (gender_source_concept_id), (gender_source_value)	Female	SNOMED	skos:closeMatch
Sex	NCIT:male	8507	Person	gender_concept_id, (gender_source_concept_id), (gender_source_value)	Male	SNOMED	skos:closeMatch
Sex	NCIT:Undetermined	4086451	Person	gender_concept_id, (gender_source_concept_id), (gender_source_value)	Patient sex unknown	SNOMED	skos:narrowMatch

Sex	NCIT:unknown	4086451	Person	gender_concept_id, (gender_source_concept_id), (gender_source_value)	Patient sex unknown	SNOMED	skos:narrowMatch
Patient status	NCIT:Subject lost to follow up	4163894	Observation		Lost to follow up	SNOMED	skos:relatedMatch
Patient status	NCIT:Refusal to participate	1314399	Observation		Patient refused to participate	HCPCS	skos:relatedMatch
Body measurement	NCIT:Body Mass Index	4245997	Measurement	measurement_source_concept_id/measurement_concept_id	Body Mass Index	SNOMED	skos:relatedMatch
Body measurement	NCIT:Height	903133	Measurement	measurement_source_concept_id/measurement_concept_id	Height	PPI	skos:relatedMatch
Body measurement	NCIT:Weight	903121	Measurement	measurement_source_concept_id/measurement_concept_id	Weight	PPI	skos:relatedMatch

Table 11: Mapping table for patient sex information, participation status and body measurement.

To do so, a mapping table was populated by every relevant conceptual entity present in the CARE-SM. Every entity type and allowed data value were defined by ontological terms, for instance, the Sex (NCIT:C28421) entity type and Female (NCIT:C16576) data value. Each ontological term in this table was then mapped to a conceptual entity present in the ATHENA vocabulary. Table 11 shows a fraction of this mapping table created, describing some of the mapped data elements.

The misalignment between similar concepts across different ontologies is one of the most complex, but also common, problems when attempting federated

exploration or data model transformation. Ontological matching has been achieved through a variety of approaches in the Semantic Web community. Some of the matching methods used are semantic similarity (an objective - automated - metric) or manual curation (a subjective metric). In this experiment we used manual curation and described the similarity between similar ontological terms using Simple Knowledge Organization System Reference (SKOS). This mapping table was reviewed by a team of healthcare semantics experts during several meetings and workshops, and the data representation and semantic similarity between concepts were unanimously approved.

Despite this concept identification table, some of the CARE-SM data elements are not found in the standardized ontologies contained within ATHENA. Although it was possible to manually map the upper-level concepts (from SIO) in CARE-SM to ATHENA, the number of domain-specific classes used by CARE-SM for specific clinical processes, anatomical structures or diagnosis was too high for manual curation.

5.5.2.3. The ETL process for converting from CARE-SM data into OMOP-CDM.

The same patient dataset generated from ERKreg during the C-PATH data model experiment was used for this study of the conversion between CARE-SM and OMOP-CDM.

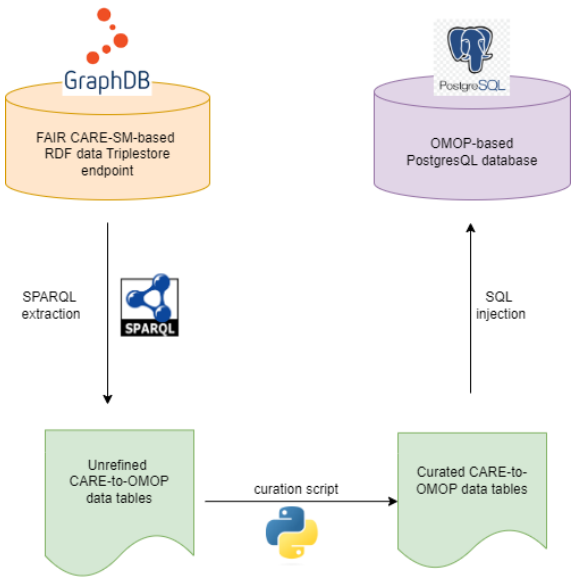


Figure 25: CARE-SM to OMOP-CDM ETL workflow.

```
PREFIX obo: <http://purl.obolibrary.org/obo/>
PREFIX sio: <http://semanticscience.org/resource/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX ordo: <http://www.orpha.net/ORDO/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT distinct ?person_id ?gender_concept_id ?year_of_birth ?month_of_birth ?day_of_birth
?birth_datetime ?race_concept_id ?ethnicity_concept_id ?location_id ?provider_id ?care_site_id
?person_source_value
 ?gender_source_value ?gender_source_concept_id ?race_source_value ?race_source_concept_id
?ethnicity_source_value ?ethnicity_source_concept_id

WHERE {
 GRAPH ?dob_g {
 ?dob_role sio:SIO_000356 ?dob_process ; a sio:SIO_000016 .
 ?dob_process a sio:SIO_000006; sio:SIO_000229 ?dob_output .
 ?dob_output sio:SIO_000628 ?dob_attribute; a sio:SIO_000015.
 ?dob_attribute a sio:SIO_000614, obo:NCIT_C68615.
 ?dob_output sio:SIO_000300 ?birth_datetime.
 }
 ?dob_g a obo:NCIT_C62143 ; sio:SIO_000068 ?dob_timeline, ?dob_event .
 ?dob_timeline a obo:NCIT_C54576, sio:SIO_000417; sio:SIO_000332 ?dob_individual .
 ?dob_individual a sio:SIO_000498 ; sio:SIO_000671 ?dob_individual_identifier .
 ?dob_individual_identifier a sio:SIO_000115 ; sio:SIO_000300 ?person_id .

 GRAPH ?sex_g {
 ?sex_role sio:SIO_000356 ?sex_process ; a sio:SIO_000016 .
 ?sex_process a sio:SIO_000006; sio:SIO_000229 ?sex_output.
 ?sex_output sio:SIO_000628 ?sex_attribute; a sio:SIO_000015, obo:NCIT_C160908.
 ?sex_attribute a sio:SIO_000614,?gender_source_value .
 FILTER (?gender_source_value != sio:SIO_000614)
 }
 ?sex_g a obo:NCIT_C62143 ; sio:SIO_000068 ?sex_timeline, ?sex_event.
 ?sex_timeline a obo:NCIT_C54576, sio:SIO_000417; sio:SIO_000332 ?sex_individual .
 ?sex_individual a sio:SIO_000498 ; sio:SIO_000671 ?sex_individual_identifier .
 ?sex_individual_identifier a sio:SIO_000115 ; sio:SIO_000300 ?person_id, ?person_source_value .
}
```

Figure 26: Exemplar SPARQL query for retrieving patient information for populating OMOP-CDM personal data table.

As it's shown in Figure 25, data extraction is performed via SPARQL querying the CARE-SM data, selecting relevant information into a resulting table partially compatible with OMOP-CDM tables. Distinct SPARQL queries (Figure 26) were created for retrieving patient data, and then populating the different OMOP-CDM data tables, according to its table documentation. Source patient information coming from CARE-SM - annotated by the use of NCIT, Orphanet, HP - is added in OMOP-CDM data table associated with the columns of "_source_concept_id". For instance, in table 12, Sex information using NCIT is populated inside the "gender_source_concept_id" column field.

CDM Field	User Guide	Datatype	Required
person_id	It is assumed that every person with a different unique identifier is in fact a different person and should be treated independently.	integer	Yes
gender_concept_id	This field is meant to capture the biological sex at birth of the Person. This field should not be used to study gender identity issues.	integer	Yes
year_of_birth	Compute age using year_of_birth.	integer	Yes
month_of_birth		integer	No
day_of_birth		integer	No
birth_datetime		datetime	No
race_concept_id	This field captures the race or ethnic background of the person.	integer	Yes
ethnicity_concept_id	This field captures Ethnicity as defined by the Office of Management and Budget (OMB) of the US Government: it distinguishes only between "Hispanic" and "Not Hispanic". Races and ethnic backgrounds are not stored here.	integer	Yes
location_id	The location refers to the physical address of the person. This field should capture the last known location of the person.	integer	No
provider_id	The Provider refers to the last known primary care provider (General Practitioner).	integer	No
care_site_id	The Care Site refers to where the Provider typically provides the primary care.	integer	No
person_source_value	Use this field to link back to persons in the source data. This is typically used for error checking of ETL logic.	varchar(50)	No
gender_source_value	This field is used to store the biological sex of the person from the source data. It is not intended for use in standard analytics but for reference only.	varchar(50)	No

gender_source_concept_id	Due to the small number of options, this tends to be zero.	integer	No
race_source_value	This field is used to store the race of the person from the source data. It is not intended for use in standard analytics but for reference only.	varchar(50)	No
race_source_concept_id	Due to the small number of options, this tends to be zero.	integer	No
ethnicity_source_value	This field is used to store the ethnicity of the person from the source data. It is not intended for use in standard analytics but for reference only.	varchar(50)	No
ethnicity_source_concept_id	Due to the small number of options, this tends to be zero.	integer	No

Table 12: OMOP-CDM data table documentation for personal patient information.

After extraction, curation procedures were performed on the output tabular results in order to derive a spreadsheet compatible with OMOP-CDM's import process. The curatorial changes are as follows:

1. Temporal data notation differs between the CARE-SM and OMOP-CDM data architectures. To ensure data consistency the curatorial change adapts values to the ISO-8601-formatted **date** (i.e. date only) to **datetime** notation. Additionally, from the patient birth date information table, this date is parsed to individual years, months, and days columns, as is shown in Table 12.
2. Several pieces of conceptual information that are mandatory for OMOP-CDM are not directly represented in CARE-SM. Curation allows the inclusion of novel data facets such as the type of clinical visit, or inpatient/outpatient status, as required by OMOP-CDM.
3. During curation, the OMOP-CDM standardized vocabulary is used whenever possible. By using the mapping table between CARE-SM and OMOP-CDM, all relevant terms identified in this table are added into OMOP-CDM data tables in the "_concept_id" column field. For example, SNOMED-CT code for "Female" from NCIT, is added in the "gender_concept_id, as it's shown in Table 12.

Once the CARE-SM-to-OMOP data tables have been curated and post-processed, they are ready for injection via SQL into the PostgreSQL relational databases.

During this step, ATC annotation, which is a non-standard terminology for OMOP-CDM was converted to the standard vocabulary of RxNorm.

In collaboration with Rodwy De Groot, PhD student and OMOP-CDM expert from Amsterdam UMC, this data was tested by the OHDSI interface called Data Quality Dashboard (DQD). DQD is deployed by a OHDSI R package called DataQualityDashboard[25].

After curating and post-processing CARE-SM-to-OMOP data tables, these datatables were injected into the PostgreSQL DQD relational database. DQD - controlled by another OHDSI R package named DatabaseConnector[26] - sets up database connections (server address and authorization) and data extraction from PostgreSQL to the DQD interface. By invoking patient data from the database in DQD, the compatibility with OMOP-CDM is evaluated, generating a report in a multidimensional description as an outcome.

EJPRD CARE-SM TO OMOP CDM

DataQualityDashboard Version: 2.6.0
Results generated at 2024-05-08 16:51:47 in 2 mins

	Verification				Validation				Total			
	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass
Plausibility	239	32	271	88%	0	0	0	-	239	32	271	88%
Conformance	725	3	728	100%	106	0	106	100%	831	3	834	100%
Completeness	390	6	396	98%	17	0	17	100%	407	6	413	99%
Total	1354	41	1395	97%	123	0	123	100%	1477	41	1518	**97%**

644 out of 1477 passed checks are Not Applicable, due to empty tables or fields.
3 out of 41 failed checks are SQL errors.
Corrected pass percentage for NA and Errors: 96% (833/871).

Figure 27: Report from Data Quality Dashboard.

DQD parameter name	DQD parameter description	DQD explanation feedback for CARE-to-OMOP information
Plausibility	Checks that data within its expected ranges.	Some observation dates occurred after the patient date of death.
Conformance	Ensures data follows the CDM	No feedback, all information

[25] https://ohdsi.github.io/DataQualityDashboard/articles/DataQualityDashboard.html
[26] https://cran.r-project.org/web/packages/DatabaseConnector/index.html

	conventions. For instance: field length and data types.	passed this parameter.
Completeness	Verifies that there are no missing or NULL values	Standardized OMOP-CDM vocabulary is missing

Table 13: Summary of DQD evaluation parameters.

DQD evaluation shows (Figure 27) a high percentage of compatible observations that passed the DQD threshold. Negative feedback retrieved from the quality assessment reported low date-related description plausibility (inconsistency in timeline of events) and the lack of standardized vocabulary for some of the observations evaluated. A summary of this DQD report is described in Table 13 and Figure 27. The whole DQD report description is in Annexes D.

6. Discussion

6.1. Improvement of CARE-SM over CDE-SM

In comparison with CDE-SM, the consistency of CARE-SM sub-components and tools required for data implementation is higher. Having a single CSV template means the data provider does not have to create a distinct export routine for each data element. Moreover, this allowed the consolidation of the numerous CDE-SM YARRRML templates into a single template, enabling easier maintenance and evolution.

Some of the advantages of CARE-SM related to the representation of longitudinal patient data were not tested due to the infrastructure limitations. Most of the ETL implementation process was deployed by the use of Fiab. This interface is snapshot-focused, unabling real-time serialization of RDF patient data. Fiab overwrites the entire FAIR data transformation for every execution, avoiding performing incremental updates. CARE-SM, by defining the timeline of event description and an event identifier for interrelating data elements, can be used for describing not only patient registry data elements, but clinical encounters for medical centers and real-timing FAIR data serialization.

The data model consistency achieved by reusing a single design pattern simplifies query, where the primary difference between data elements are the ontological classes that define the various sub-elements of a data type. Thus, through minor adjustments to an overall SPARQL query template, any of the CARE-SM data elements can be explored in the same way. An example of this unified SPARQL query is in Annexes C.

This improvement is demonstrated through our implementation of Beacon-2 API for CARE-SM. SPARQL queries are built by a combination of SPARQL "fragments" or "blocks". Most of these SPARQL fragments are reused in every data element filter. This is an advantage for dynamically creating complex filters in a single Beacon request. CDE-SM aggregated representation possibilities the inclusion of partially filled RDF data representation, making confidence impossible for Beacon responses.

6.2. Demonstration of Interoperability between sites that implement CARE-SM

The automated deployment of the Beacon-2 API over the numerous participants in EJP-RD is a clear demonstration of the achievement of interoperability through the use of CARE-SM. In that respect, this study was successful. However, this success is not unexpected - it is generally true that sites that use the same structure and the same vocabularies will be interoperable, whether they use FAIR/semantic technologies or not. Nevertheless, it is rare for non-coordinating sites to adopt the same models and terminologies, so by simplifying the achievement of a common framework via shared templates and the straightforward creation of CSV files, it is demonstrably plausible to pursue large-scale interoperability using this approach. The limitations are discussed further in a subsequent section.

6.3. Interoperability with other standards

We have explored the question of whether the semantic agents we have developed are capable of effectively handling schema integration, or if the future of the Semantic Web is destined to rely on non-automated processes and human

perception. If the semantics required for interoperability cannot be derived from the schema, the question of where else they can originate, arises.

Interoperability - the "I" in FAIR - is highlighted as one of the greatest challenges in data science. Historically, schema mapping cannot be automated for several reasons: the available information, in the form of metadata and entity-relationship representations from schemas is, in almost all cases, insufficient for current agents to automate data mappings between non-coordinating schemas. Moreover, the lack of consensus in the concept representation across ontologies limits automated mapping across different vocabularies.

A large and well-funded international consortium - EJP-RD - has made efforts to address this challenge, and embarked on several schema mapping investigations, two of which are described in this thesis. However, it took a considerable amount of time and resources. While we have made notable progress towards interoperability within the consortium by creating a stable semantic model for rare diseases (e.g. Beacon-2 over CARE-SM), the problem of insufficient automation of query between EJP and third-party data remains prevalent beyond the end of the EJP-RD project. Mapping activities alone have not achieved full interoperability between (even) semantic-based models, and in fact, this thesis contributes additional information about why these attempts so frequently fail.

Mapping independent datasets from two distinct data models to an upper-level schema, such as Biolink, offered only a potential starting point for interoperability. These shared concepts could be used as "anchor points" between the distinct representations, such that queries could use these to interrogate the same data from both models, as shown in Table 10.

Nevertheless, those queries also reveal the significant limitations of this approach to interoperability. Several actions taken during this study would not be feasible in a real scenario, or at-scale. For instance, adding additional new ontological classes into any pre-existing data would not be possible unless you were the owner of that data. It might be possible to create an independent dataset containing the mapping information (i.e. the information in table 9) that is then accessed dynamically at the time of query; however, this is far from trivial, and requires extensive manual intervention. It is not how we would envision "interoperability" to be achieved in any meaningful way.

Another consideration arises from adding Biolink classes directly into the data: Biolink is a schema, not an ontology. The logical meanings of "type" in a schema are distinct from the rigorous definition of "type" in an OWL Ontology. The C-Path data model is based on an OWL representation of Biolink, so adding this mapping directly into their data will do little harm. However, generally adding schema concepts as classes in the data models will break the logic of an OWL-based model, making them unusable for the purpose of logical reasoning. Other limitations are similarly difficult to resolve. For example:

- Disparate treatment of time information, such as age versus birth date, cannot be resolved in an automated way, even with schema mapping, since this can only be mitigated via a calculation based on some prior knowledge about the datasets.

- Inconsistent use of units across datasets results in data that cannot be immediately integrated. If the measurement unit is captured in the dataset, it has been shown that data integration can be automated (Samadian et al., 2014), however this is not always the case, and requires dedicated infrastructure that is independent of typical federated query tools.

- There continues to be a lack of standardized vocabularies for many domains of clinical data, and thus there may be no way to achieve a formal mapping via an external "anchor" terminology.

Another effort from EJP-RD to achieve data interoperability across healthcare standards was mapping CDE-SM/CARE-SM to OMOP-CDM. Although ETL workflow showed the feasibility of creating OMOP-CDM compatible data tables, the success was limited.

None of the source annotations from the CARE-SM are accepted by ATHENA. This experiment was limited by the lack of third-parties mapping across the most frequent ontologies utilized in CARE-SM (NCIT, HP, Orphanet) and SNOMED-CT, LOINC or RxNorm. Only manual curation for well-known permitted values - such as patient sex or participant status - and third-party ATC to RxNorm were

possible. Only a small fraction of NCIT codes is represented in the ATHENA vocabulary, however none of them were applicable to our case.

By these limited mappings across ontologies, a low number of data elements are successfully described with standardized vocabulary of ATHENA. Without this standardized vocabulary, OHDSI interfaces cannot facilitate data discoverability or analyses to the user, limiting any capability of most of the OHDSI frameworks available.

Lack of a standardized vocabulary is not the only limitation identified during this data model conversion. Although the data is described by a standardized vocabulary, it falls short of a sufficient semantic layer of metadata. Much of their model documentation is based on narrative descriptions that are not interoperable. Lexical descriptions, as the only way to interpret OMOP-CDM, is not sufficient in some cases to distinguish the destination of the data element across all the different data tables this model defines.

Overall, it seems clear that the achievement of interoperability between our semantic model and third-party semantic models was limited, at best, and far from automatable.

6.4. FAIR is necessary but not sufficient

None of the accomplishments in our creation of a FAIR data model enhanced significantly the data interoperability with other studied standards. Most of the advantages with CDE-SM, and the subsequent CARE-SM are at a structural or ease-of-deployment level.
FAIR data representation alone, therefore, seems not to be sufficient for achieving data interoperability across non-coordinating resources. Interoperability is achieved only through a combination of FAIR, bringing unique identifiers and semantics into the data to ensure consistency in interpretation, and shared models, allowing the same exploratory tools to be reused over all participating sites. This is important information for those embarking on their own FAIRification initiatives - it should temper expectations, and should encourage the exploration of shared models within the community in which you wish to achieve federation.

6.5. Future directions

While CARE-SM was designed for representing patient clinical data, the overall model is grounded in the upper-level SIO ontology, which is a general model for entity-relations in all domains of science. As such, we believe that CARE-SM could be adapted for other purposes such as biobanking, by reusing nodes such as **identifier, process, output,** and **attribute** or **quality**. The **specification** node, in this case, would represent the sampling protocol. This reusability is currently being tested by other projects, for example, a project focused on seed-banks, where the core CARE-SM model is being applied to the collection of samples in field studies.

Another future direction, not addressed in the scope of this thesis, is the use of Description Logic (DL) (Rudolph, 2011) reasoners to interpret data. Significant effort was dedicated to avoiding logical inconsistencies in the CARE-SM model, such that model-compliant data could be explored and automatically classified by a reasoner. By ensuring the logical consistency of the model itself, it should be possible to then leverage the logical features of the third-party ontologies used by the model. For example, we should be able to automatically identify all data that falls within the subClass hierarchy of a phenotypic characteristic from the HPO. Given that projects such as the EJP-RD are requiring data to be represented using CARE-SM, this kind of automated query expansion and logical classification should be a high priority for future studies.

7. Conclusions

- CARE-SM represents rare disease patient data registries in a FAIR manner that scales to represent not only the set of common data elements enumerated by the European Commission but a much wider range of healthcare data types.

- The consistency in CARE-SM representation of diverse data types enables a high degree of federated data exploration between sites that have adopted this model, even when the underlying data is from distinct sub-domains of clinical investigation.

- FAIR-compliance is not sufficient for achieving data interoperability. A shared model is also required.

- FAIR-compliance is not sufficient to enhance interoperability with third-party standards. Manual mapping continues to be the only approach for reliable data integration.

8. References

Alarcon, P., Braun, I., Hartley, E., Olson, D., Benis, N., Cornet, R., Wilkinson, M., & Walls, R. L. (2023). Leveraging Biolink as a "Rosetta Stone" Between C-Path and EJP-RD Semantic Models Provides Emergent Interoperability. *Journal of the Society for Clinical Data Management*, *3*(1), Article 1. https://doi.org/10.47912/jscdm.130

Altman, R. B., Buda, M., Chai, X. J., Carillo, M. W., Chen, R. O., & Abernethy, N. F. (1999). RiboWeb: An ontology-based system for collaborative molecular biology. *IEEE Intelligent Systems*, *14*(5), 68–76. https://doi.org/10.1109/5254.796092

Bassanese, G., Wlodkowski, T., Servais, A., Heidet, L., Roccatello, D., Emma, F., Levtchenko, E., Ariceta, G., Bacchetta, J., Capasso, G., Jankauskiene, A., Miglinas, M., Ferraro, P. M., Montini, G., Oh, J., Decramer, S., Levart, T. K., Wetzels, J., Cornelissen, E., … Schaefer, F. (2021). The European Rare Kidney Disease Registry (ERKReg): Objectives, design and initial results. *Orphanet Journal of Rare Diseases*, *16*(1), Article 1. https://doi.org/10.1186/s13023-021-01872-8

Bayat, A. (2002). Science, medicine, and the future: Bioinformatics. *BMJ*, *324*(7344), 1018–1022. https://doi.org/10.1136/bmj.324.7344.1018

Beno, M., Filtz, E., Kirrane, S., & Polleres, A. (2019). Doc2RDFa: Semantic Annotation for Web Documents. In *Proceedings of the Posters and Demos Track of the 15th International Conference on Semantic Systems (SEMANTiCS 2019)*. https://research.wu.ac.at/en/publications/doc2rdfa-semantic-annotation-for-web-documents-4

BergerBonnie, M, D., & William, Y. (2016). Computational biology in the 21st century. *Communications of the ACM*. https://doi.org/10.1145/2957324

Briney, K. A. (2018). The Problem with Dates: Applying ISO 8601 to Research Data Management. *Journal of eScience Librarianship*, *7*(2), Article 2. https://doi.org/10.7191/jeslib.2018.1147

Bueno-de-la-Fuente, G. (2008). *The Simple Knowledge Organization System (SKOS): A situation report for the HIVE Project*. https://hdl.handle.net/10016/9090

Ceusters, W., Smith, B., & Goldberg, L. (2018). A Terminological and Ontological Analysis of the NCI Thesaurus. *Methods of Information in*

Medicine, *44*, 498–507. https://doi.org/10.1055/s-0038-1634000

Damme, P. van, Moreno, P. A., Bernabé, C. H., Ballesteros, A. C., Cornec, C. M. A. L., Vieira, B. D. S., Velde, K. J. van der, Zhang, S., Carta, C., Cornet, R., Hoen, P. A. C. 't, Jacobsen, A., Swertz, M. A., Roos, M., & Benis, N. (2023). *A Resource for Guiding Data Stewards to Make European Rare Disease Patient Registries FAIR* (1). *22*(1), Article 1. https://doi.org/10.5334/dsj-2023-012

Dumontier, M., Baker, C. J., Baran, J., Callahan, A., Chepelev, L., Cruz-Toledo, J., Del Rio, N. R., Duck, G., Furlong, L. I., Keath, N., Klassen, D., McCusker, J. P., Queralt-Rosinach, N., Samwald, M., Villanueva-Rosales, N., Wilkinson, M. D., & Hoehndorf, R. (2014). The Semanticscience Integrated Ontology (SIO) for biomedical research and knowledge discovery. *Journal of Biomedical Semantics*, *5*(1), 14. https://doi.org/10.1186/2041-1480-5-14

Dunnen, J. T. den, Dalgleish, R., Maglott, D. R., Hart, R. K., Greenblatt, M. S., McGowan-Jordan, J., Roux, A.-F., Smith, T., Antonarakis, S. E., & Taschner, P. E. M. (2016). HGVS Recommendations for the Description of Sequence Variants: 2016 Update. *Human Mutation*, *37*(6), 564–569. https://doi.org/10.1002/humu.22981

El-Sappagh, S., Franda, F., Ali, F., & Kwak, K.-S. (2018). SNOMED CT standard ontology based on the ontology for general medical science. *BMC Medical Informatics and Decision Making*, *18*(1), 76. https://doi.org/10.1186/s12911-018-0651-5

European Commission. Directorate General for Research and Innovation.

& PwC EU Services. (2018). *Cost-benefit analysis for FAIR research data: Cost of not having FAIR research data.* Publications Office. https://data.europa.eu/doi/10.2777/02999

Fragoso, G., Coronado, S. de, Haber, M., Hartel, F., & Wright, L. (2004). Overview and utilization of the NCI Thesaurus. *Comparative and Functional Genomics*, *5*(8), 648–654. https://doi.org/10.1002/cfg.445

Gómez-López, G., Dopazo, J., Cigudosa, J. C., Valencia, A., & Al-Shahrour, F. (2019). Precision medicine needs pioneering clinical bioinformaticians. *Briefings in Bioinformatics*, *20*(3), 752–766. https://doi.org/10.1093/bib/bbx144

Hamosh, A., Scott, A. F., Amberger, J., Valle, D., & McKusick, V. A. (2000). Online Mendelian Inheritance In Man (OMIM). *Human Mutation*, *15*(1), 57–61. https://doi.org/10.1002/(SICI)1098-1004(200001)15:1<57::AID-HUMU12>3.0.CO;2-G

Harrow, I. (2019). Ontology mapping for semantically enabled applications. *Drug Discovery Today*, *24*(10), 2068–2075. https://doi.org/10.1016/j.drudis.2019.05.020

Hogeweg, P. (2011). The Roots of Bioinformatics in Theoretical Biology. *PLoS Computational Biology*, *7*(3), e1002021. https://doi.org/10.1371/journal.pcbi.1002021

Hommeaux, E., & Seaborne. (2008). *SPARQL query language for RDF, W3C Recommendation | Request PDF*. ResearchGate. https://www.researchgate.net/publication/272353260_SPARQL_query_language_for_RDF_W3C_Recommendation

Horrocks, I. (2002). DAML+OIL: A Description Logic for the Semantic Web. *IEEE Data Eng. Bull.*, *25*(1), 4–9.

Hripcsak, G., Duke, J. D., Shah, N. H., Reich, C. G., Huser, V., Schuemie, M. J., Suchard, M. A., Park, R. W., Wong, I. C. K., Rijnbeek, P. R., van der Lei, J., Pratt, N., Nor&#233, N, G. N., Li, Y.-C., Stang, P. E., Madigan, D., & Ryan, P. B. (2015). Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. In *MEDINFO 2015: eHealth-enabled Health* (pp. 574–578). IOS Press. https://doi.org/10.3233/978-1-61499-564-7-574

I, R., M, Z., C, R., M, S., & F, B. (2021). The Usage of OHDSI OMOP - A Scoping Review. *Studies in Health Technology and Informatics*, *283*. https://doi.org/10.3233/SHTI210546

Iglesias, E., Jozashoori, S., Chaves-Fraga, D., Collarana, D., & Vidal, M.-E. (2020). SDM-RDFizer: An RML Interpreter for the Efficient Creation of RDF Knowledge Graphs. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 3039–3046. https://doi.org/10.1145/3340531.3412881

Jacobs, I. (n.d.). *Linking the World's Information: Essays on Tim Berners-Lee's Invention of the World Wide Web | ACM Books | ACM Digital Library* (world). ACM Books. https://doi.org/10.1145/3591366.3591380

Kaliyaperumal, R., Wilkinson, M. D., Moreno, P. A., Benis, N., Cornet, R., dos Santos Vieira, B., Dumontier, M., Bernabé, C. H., Jacobsen, A., Le Cornec, C. M. A., Godoy, M. P., Queralt-Rosinach, N., Schultze Kool, L. J., Swertz, M. A., van Damme, P., van der Velde, K. J., Lalout, N., Zhang, S.,

& Roos, M. (2022). Semantic modelling of common data elements for rare disease registries, and a prototype workflow for their deployment over registry data. *Journal of Biomedical Semantics*, *13*(1), Article 1. https://doi.org/10.1186/s13326-022-00264-6

Köhler, S., Gargano, M., Matentzoglu, N., Carmody, L. C., Lewis-Smith, D., Vasilevsky, N. A., Danis, D., Balagura, G., Baynam, G., Brower, A. M., Callahan, T. J., Chute, C. G., Est, J. L., Galer, P. D., Ganesan, S., Griese, M., Haimel, M., Pazmandi, J., Hanauer, M., … Robinson, P. N. (2021). The Human Phenotype Ontology in 2021. *Nucleic Acids Research*, *49*(D1), D1207–D1217. https://doi.org/10.1093/nar/gkaa1043

Lebo, T., Sahoo, S., McGuinness, D., Belhajjame, K., Cheney, J., Corsar, D., Garijo, D., Soiland-Reyes, S., Zednik, S., & Zhao, J. (2013). Prov-o: The prov ontology. *W3C Recommendation*, *30*. https://pure.manchester.ac.uk/ws/files/31956469/FULL_TEXT.PDF

Li, H., Dragisic, Z., Faria, D., Ivanova, V., Jiménez-Ruiz, E., Lambrix, P., & Pesquita, C. (2019). User validation in ontology alignment: Functional assessment and impact. *The Knowledge Engineering Review*, *34*, e15. https://doi.org/10.1017/S0269888919000080

Luscombe, N. M., Greenbaum, D., & Gerstein, M. (2018). What is Bioinformatics? A Proposed Definition and Overview of the Field. *Methods of Information in Medicine*, *40*, 346–358. https://doi.org/10.1055/s-0038-1634431

Manuja, M., & Garg, D. (2011). *Semantic Web Mining of Unstructured Data: Challenges and Opportunities | Request PDF*. ResearchGate.

https://www.researchgate.net/publication/251422283_Semantic_Web_Mining_of_Unstructured_Data_Challenges_and_Opportunities

McGuinness, D. L., & Van Harmelen, F. (2004). OWL web ontology language overview. *W3C Recommendation*, *10*(10), 2004.

Mons, B., Neylon, C., Velterop, J., Dumontier, M., da Silva Santos, L. O. B., & Wilkinson, M. D. (2017). Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. *Information Services & Use*, *37*(1), 49–56. https://doi.org/10.3233/ISU-170824

Otte, J. N., Beverley, J., & Ruttenberg, A. (2022). BFO: Basic Formal Ontology. *Applied Ontology*, *17*(1), 17–43. https://doi.org/10.3233/AO-220262

Pérez, J., ArenasMarcelo, & GutierrezClaudio. (2009). Semantics and complexity of SPARQL. *ACM Transactions on Database Systems (TODS)*. https://doi.org/10.1145/1567274.1567278

Putman, T. E., Schaper, K., Matentzoglu, N., Rubinetti, V. P., Alquaddoomi, F. S., Cox, C., Caufield, J. H., Elsarboukh, G., Gehrke, S., Hegde, H., Reese, J. T., Braun, I., Bruskiewich, R. M., Cappelletti, L., Carbon, S., Caron, A. R., Chan, L. E., Chute, C. G., Cortes, K. G., … Munoz-Torres, M. C. (2024). The Monarch Initiative in 2024: An analytic platform integrating phenotypes, genes and diseases across species. *Nucleic Acids Research*, *52*(D1), D938–D949. https://doi.org/10.1093/nar/gkad1082

Raffl, C., Hofkirchner, W., Fuchs, C., & Schafranek, M. (2008). The Web as

Techno-Social System: The Emergence of Web 3.0. *Cybernetics and Systems*, 604–609.

Rambla, J., Baudis, M., Ariosa, R., Beck, T., Fromont, L. A., Navarro, A., Paloots, R., Rueda, M., Saunders, G., Singh, B., Spalding, J. D., Törnroos, J., Vasallo, C., Veal, C. D., & Brookes, A. J. (2022). Beacon v2 and Beacon networks: A "lingua franca" for federated data discovery in biomedical genomics, and beyond. *Human Mutation*, *43*(6), 791–799. https://doi.org/10.1002/humu.24369

Reich, C., Ostropolets, A., Ryan, P., Rijnbeek, P., Schuemie, M., Davydov, A., Dymshyts, D., & Hripcsak, G. (2024). OHDSI Standardized Vocabularies—A large-scale centralized reference ontology for international data harmonization. *Journal of the American Medical Informatics Association*, *31*(3), 583–590. https://doi.org/10.1093/jamia/ocad247

Rudolph, S. (2011). Foundations of Description Logics. *Reasoning Web. Semantic Technologies for the Web of Data*, 76–136. https://doi.org/10.1007/978-3-642-23032-5_2

Samadian, S., McManus, B., & Wilkinson, M. D. (2014). Automatic detection and resolution of measurement-unit conflicts in aggregated data. *BMC Medical Genomics*, *7*(1), Article 1. https://doi.org/10.1186/1755-8794-7-S1-S12

Sunil Krishnan, G., Joshi, A., & Kaushik, V. (2021). Bioinformatics in Personalized Medicine. In *Advances in Bioinformatics* (pp. 303–315). Springer, Singapore. https://doi.org/10.1007/978-981-33-6191-1_15

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., Glusman, G., Hadlock, J. J., Harris, N. L., Joshi, A., Putman, T., Qin, G., Ramsey, S. A., Shefchek, K. A., Solbrig, H., … Consortium, T. B. D. T. (2022). Biolink Model: A universal schema for knowledge graphs in clinical, biomedical, and translational science. *Clinical and Translational Science*, *15*(8), 1848–1855. https://doi.org/10.1111/cts.13302

Vasant, D., Chanas, L., Malone, J., Hanauer, M., Olry, A., Jupp, S., Robinson, P. N., Parkinson, H., & Rath, A. (2014). Ordo: An ontology connecting rare disease, epidemiology and genetic data. *Proceedings of ISMB*, *30*. https://www.researchgate.net/profile/Drashtti-Vasant/publication/281824026_ORDO_An_Ontology_Connecting_Rare_Disease_Epidemiology_and_Genetic_Data/links/55f99bc408aeafc8ac266edf/ORDO-An-Ontology-Connecting-Rare-Disease-Epidemiology-and-Genetic-Data.pdf

Weldring, T., & Smith, S. M. S. (2013). Article Commentary: Patient-Reported Outcomes (PROs) and Patient-Reported Outcome Measures (PROMs). *Health Services Insights*, *6*, HSI.S11093. https://doi.org/10.4137/HSI.S11093

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship.

Scientific Data, *3*(1), Article 1. https://doi.org/10.1038/sdata.2016.18

Wood, F., & Guinter, T. (2008). Evolution and Implementation of the CDISC Study Data Tabulation Model (SDTM). *Pharmaceutical Programming*. https://doi.org/10.1179/175709208X334623

Zhou, L., Thiéblin, E., Cheatham, M., Faria, D., Pesquita, C., Trojahn, C., & Zamazal, O. (2020). Towards evaluating complex ontology alignments. *The Knowledge Engineering Review*, *35*, e21. https://doi.org/10.1017/S0269888920000168

Zhu, G. (2017). *Computing Semantic Similarity of Concepts in Knowledge Graphs*. https://ieeexplore.ieee.org/document/7572993

9. Annexes

Annexes A: CDEs representations

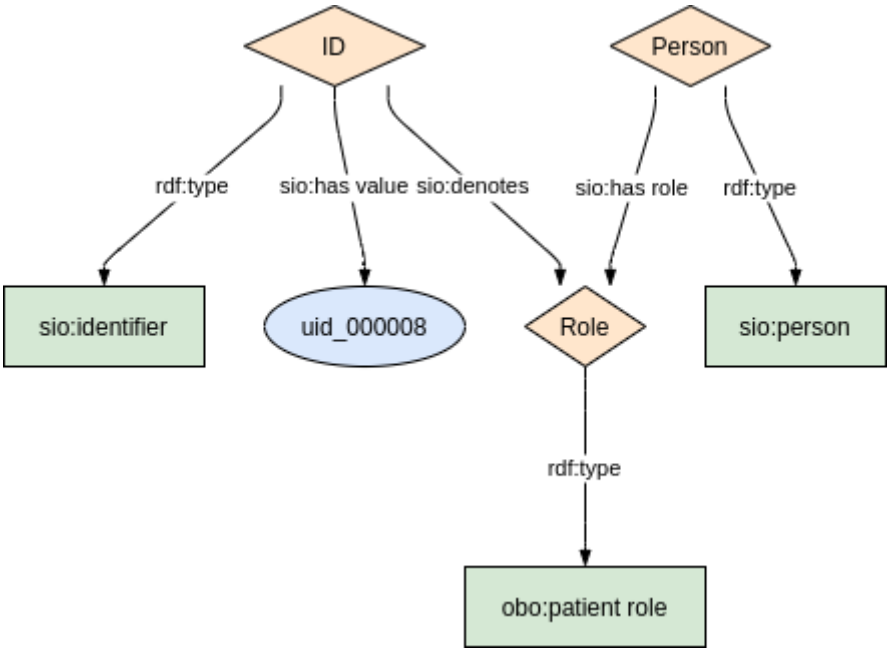


Figure A1: CDE-SM Pseudonym information.

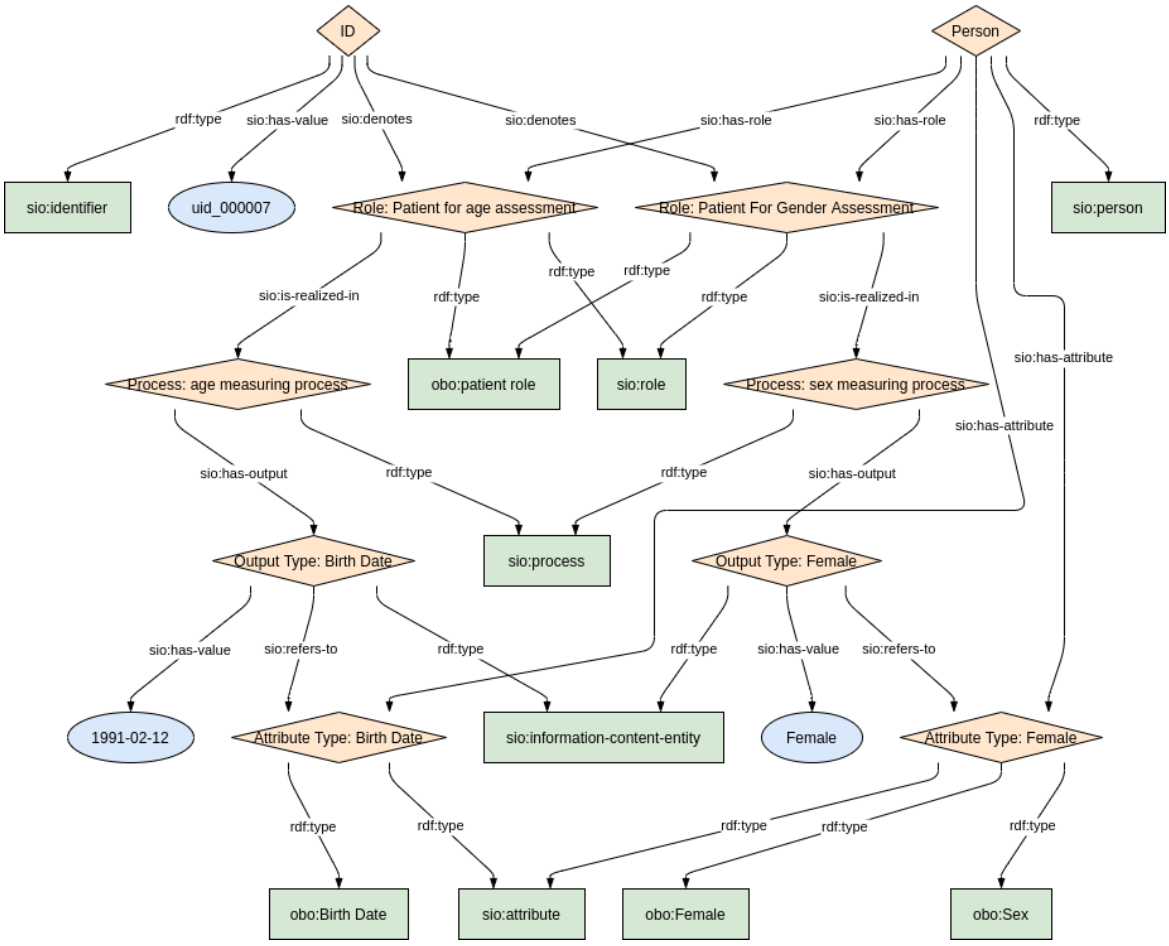


Figure A2: CDE-SM Personal information.

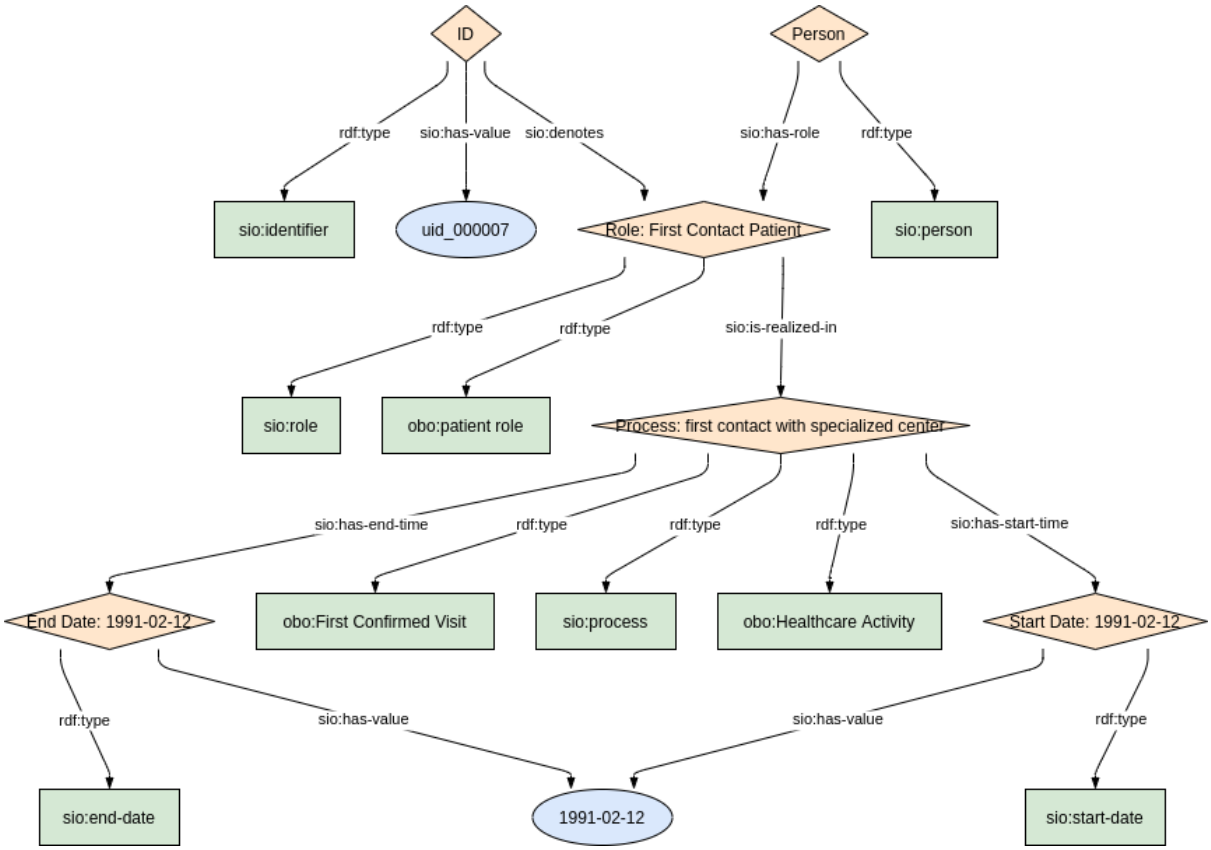


Figure A3: CDE-SM Care pathway information.

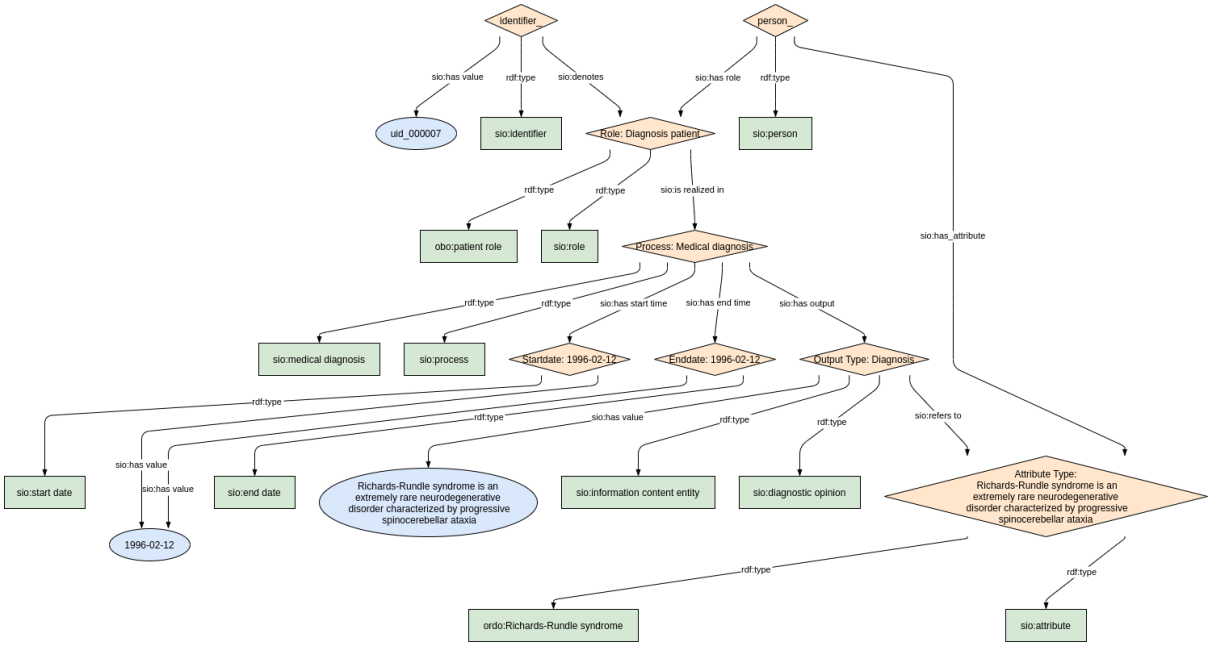


Figure A4: CDE-SM Diagnosis information

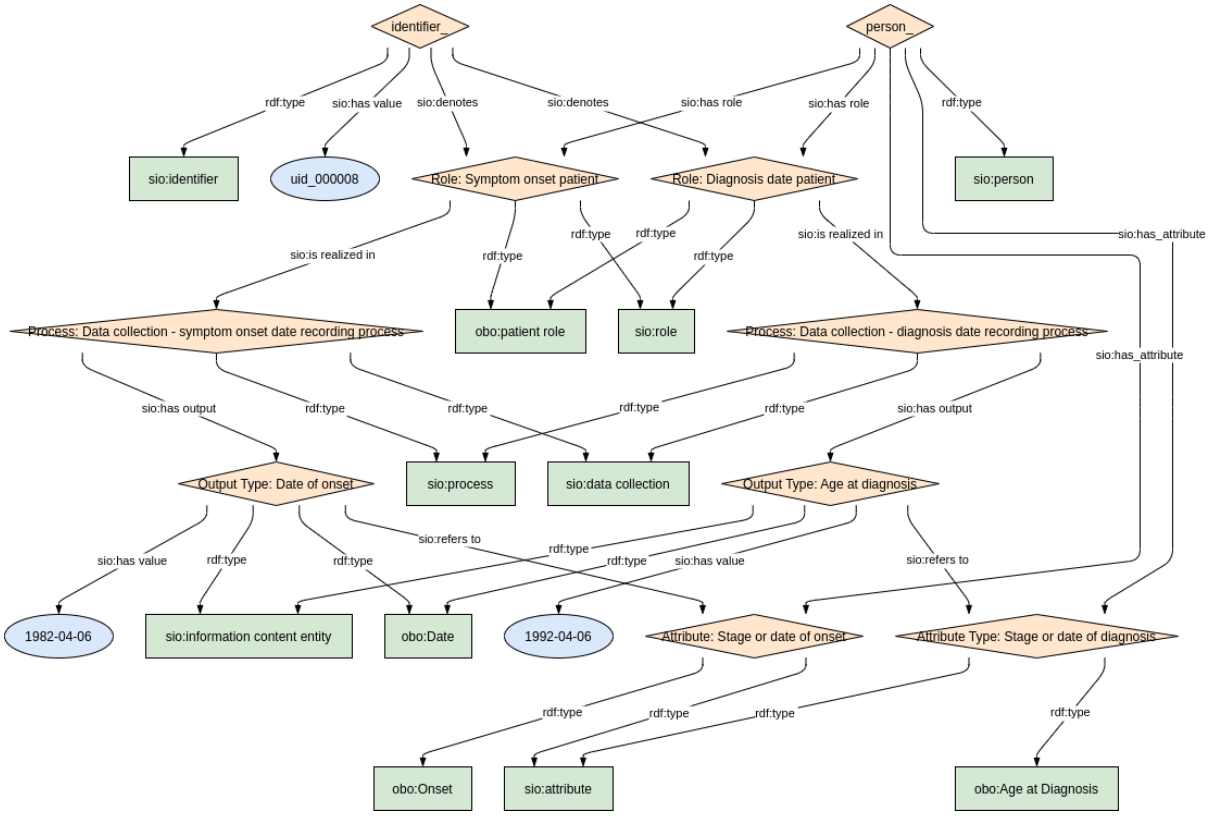


Figure A5: CDE-SM Disease history information.

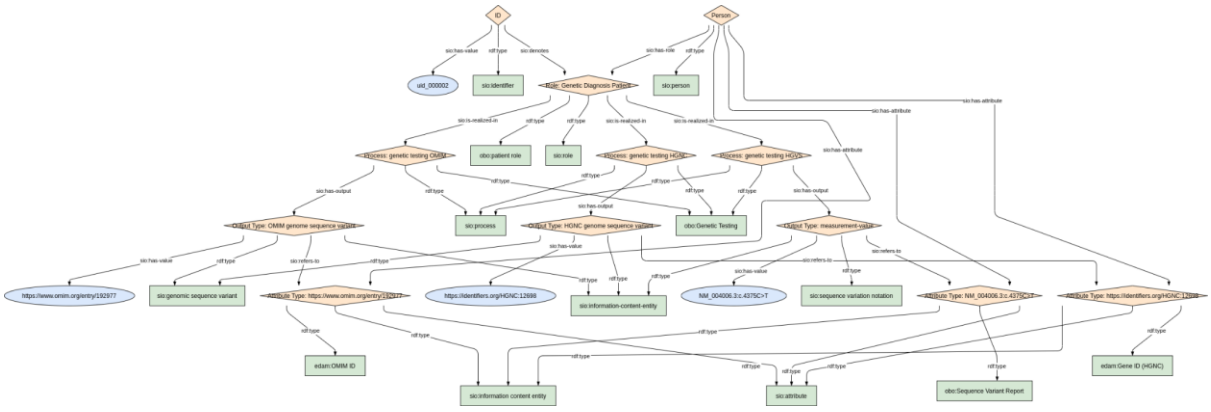


Figure A6: CDE-SM Genetic diagnosis information

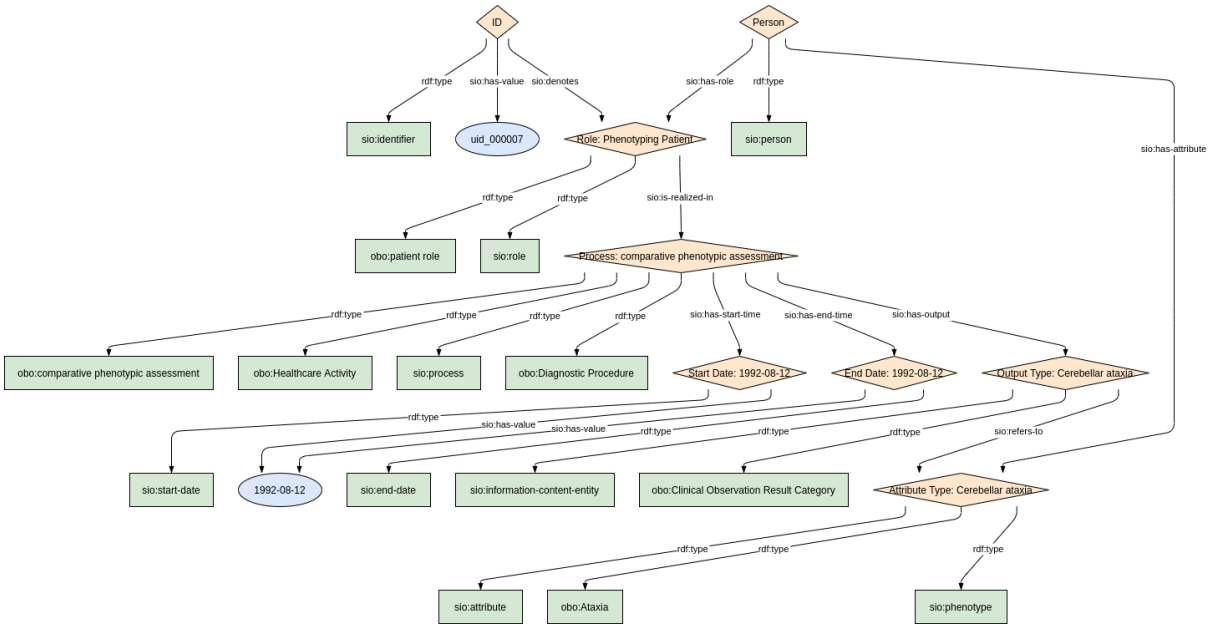


Figure A7: CDE-SM Phenotype information.

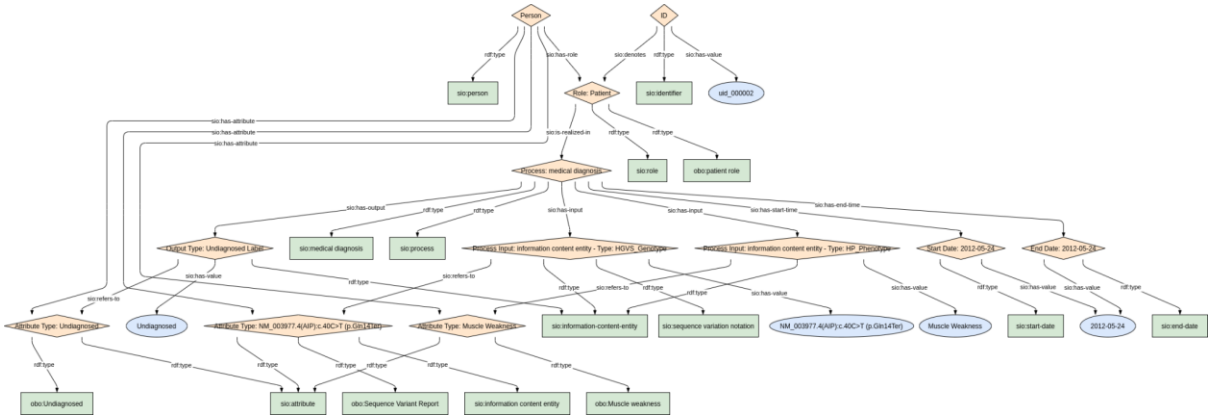


Figure A8: CDE-SM Undiagnosis information.

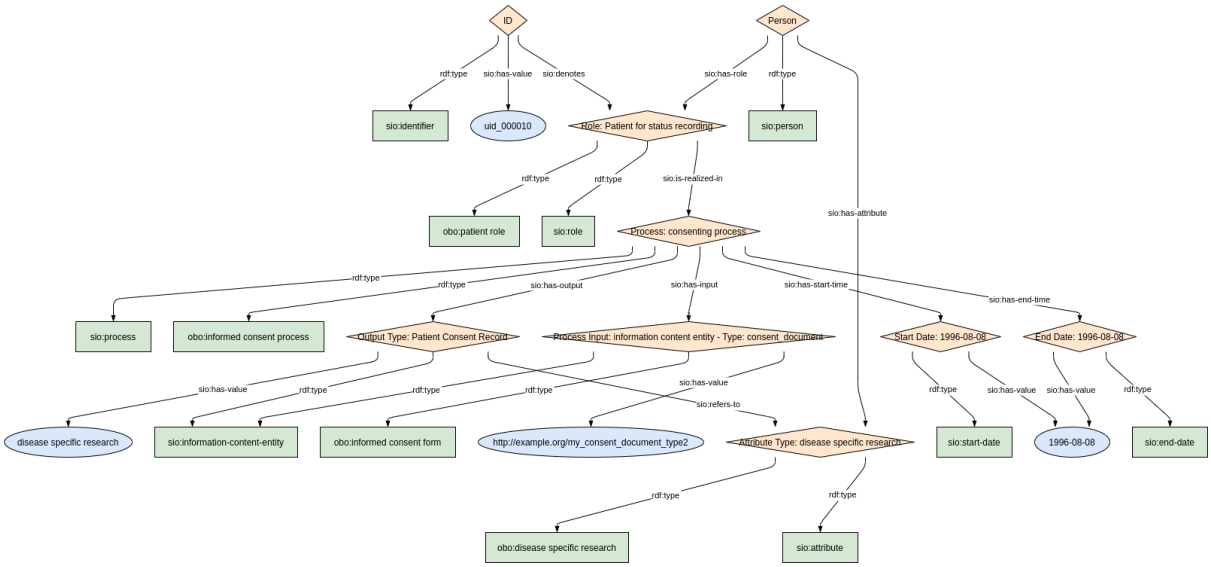


Figure A9: CDE-SM Consent information.

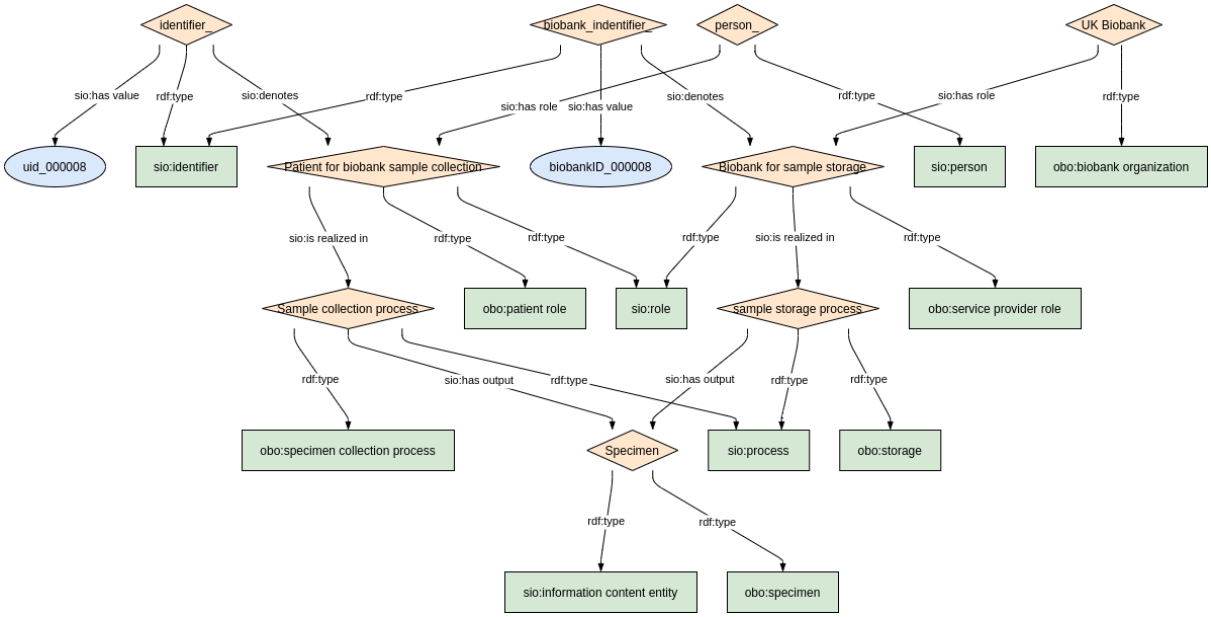


Figure A10: CDE-SM Biobank information.

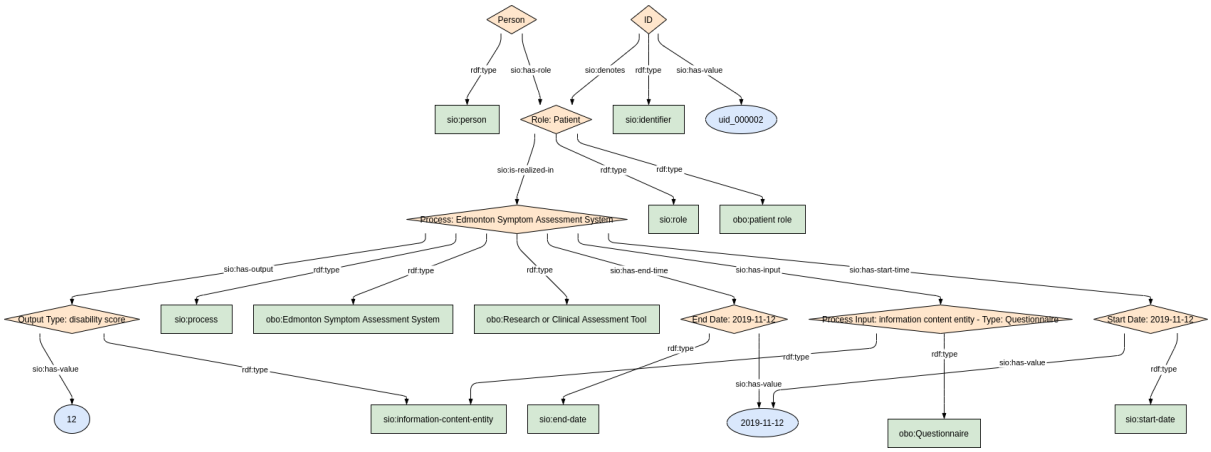


Figure A11: CDE-SM Patient disability information

Annexes B: CARE-SM representations

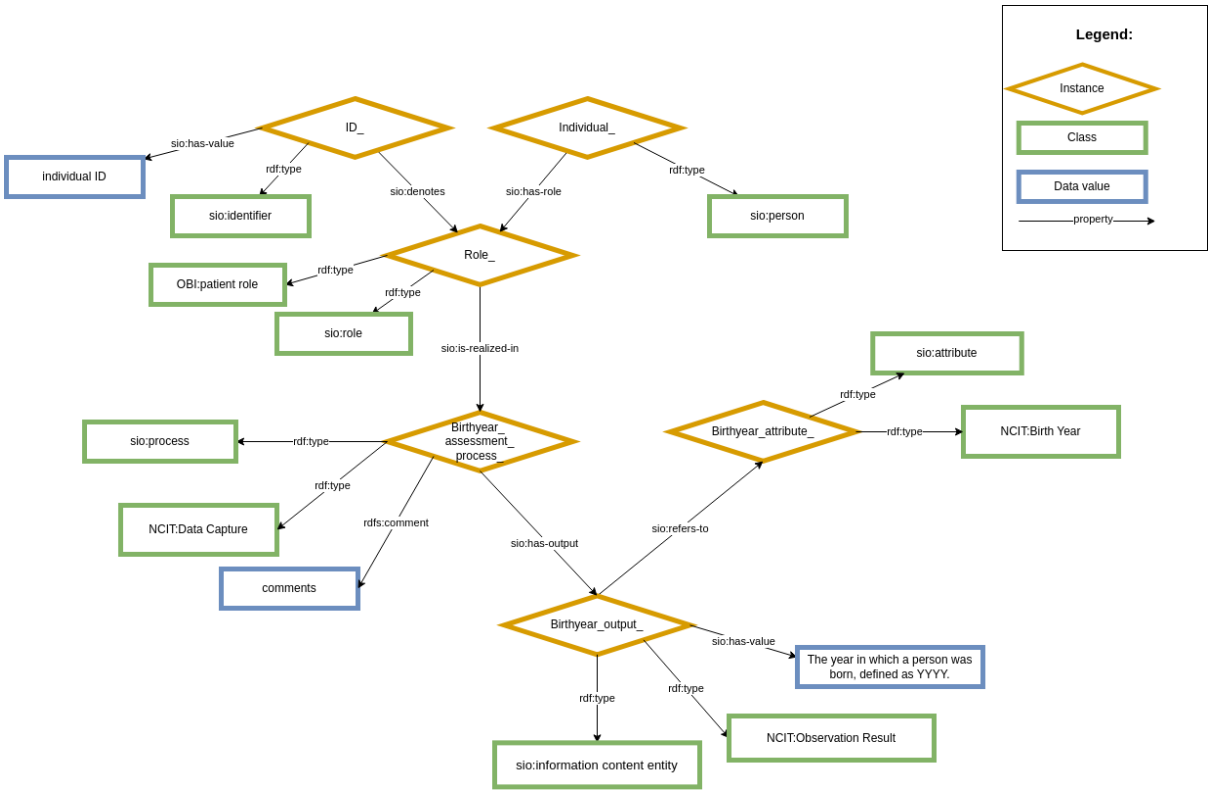


Figure B.1: CARE-SM Birthyear information.

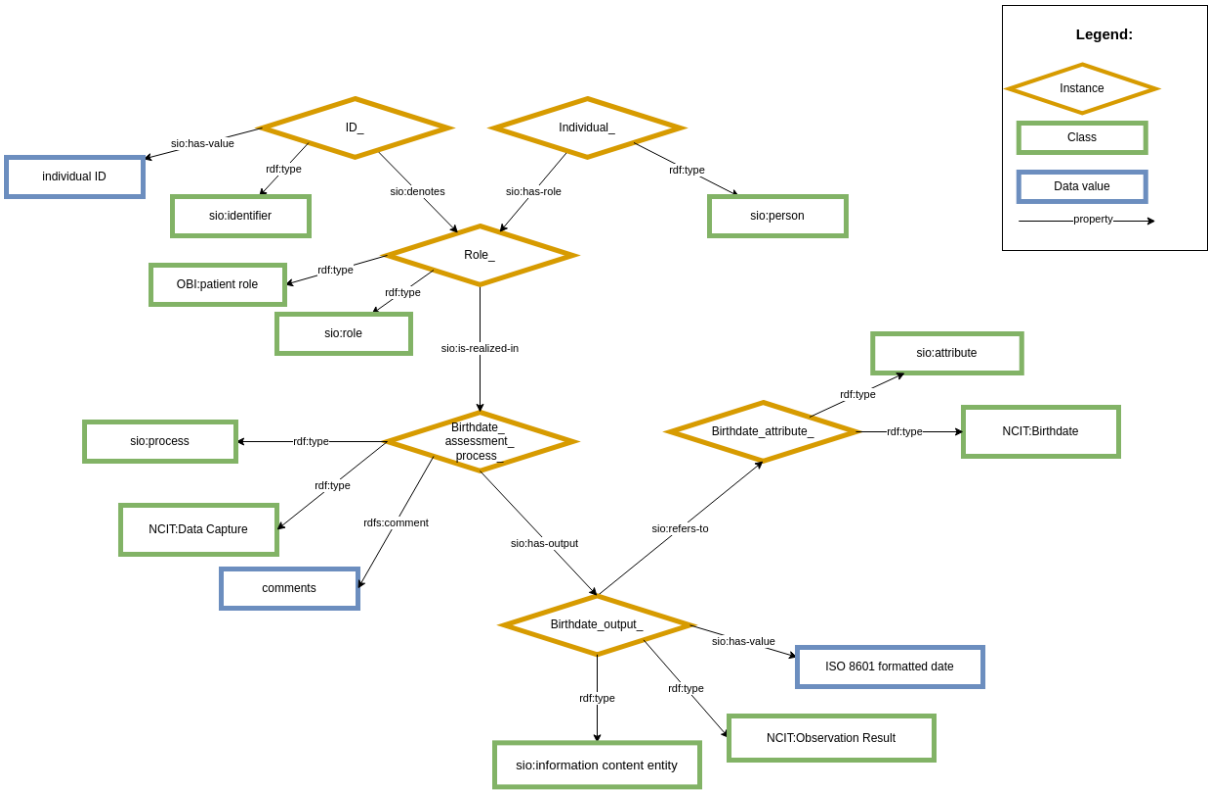


Figure B.2: CARE-SM Birthdate information.

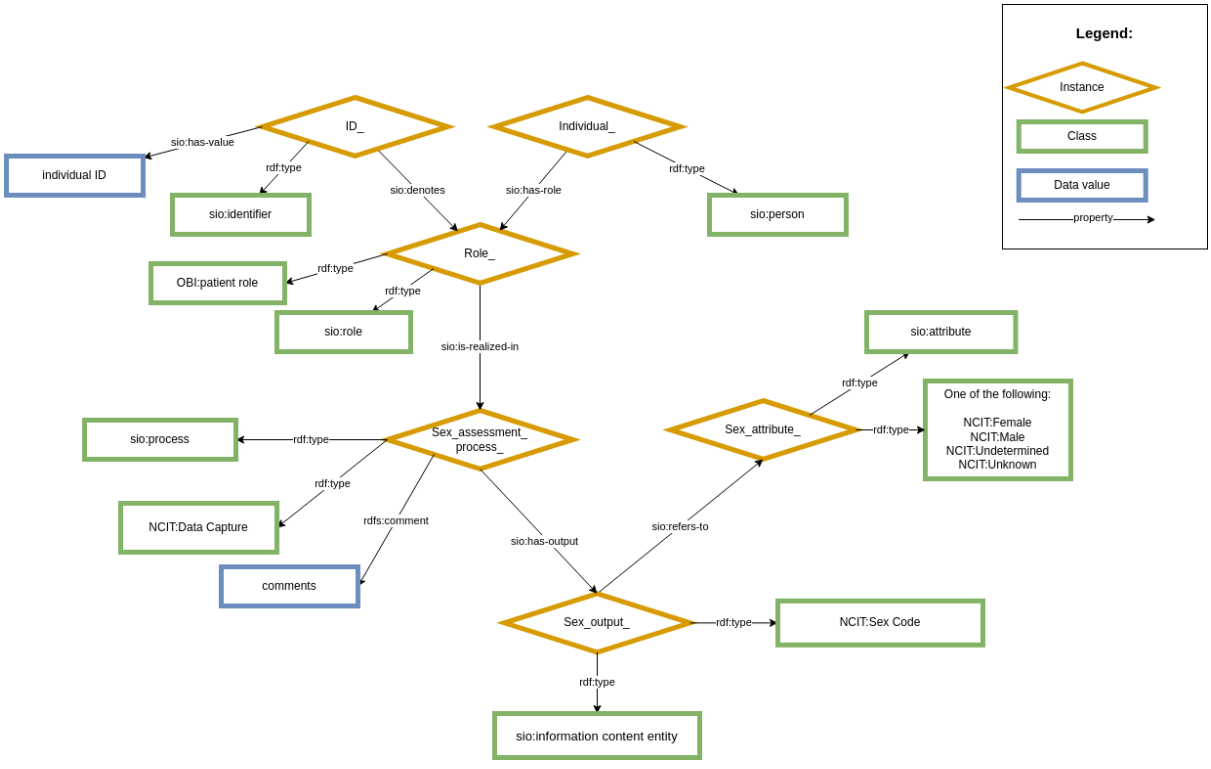


Figure B.3: CARE-SM Sex information.

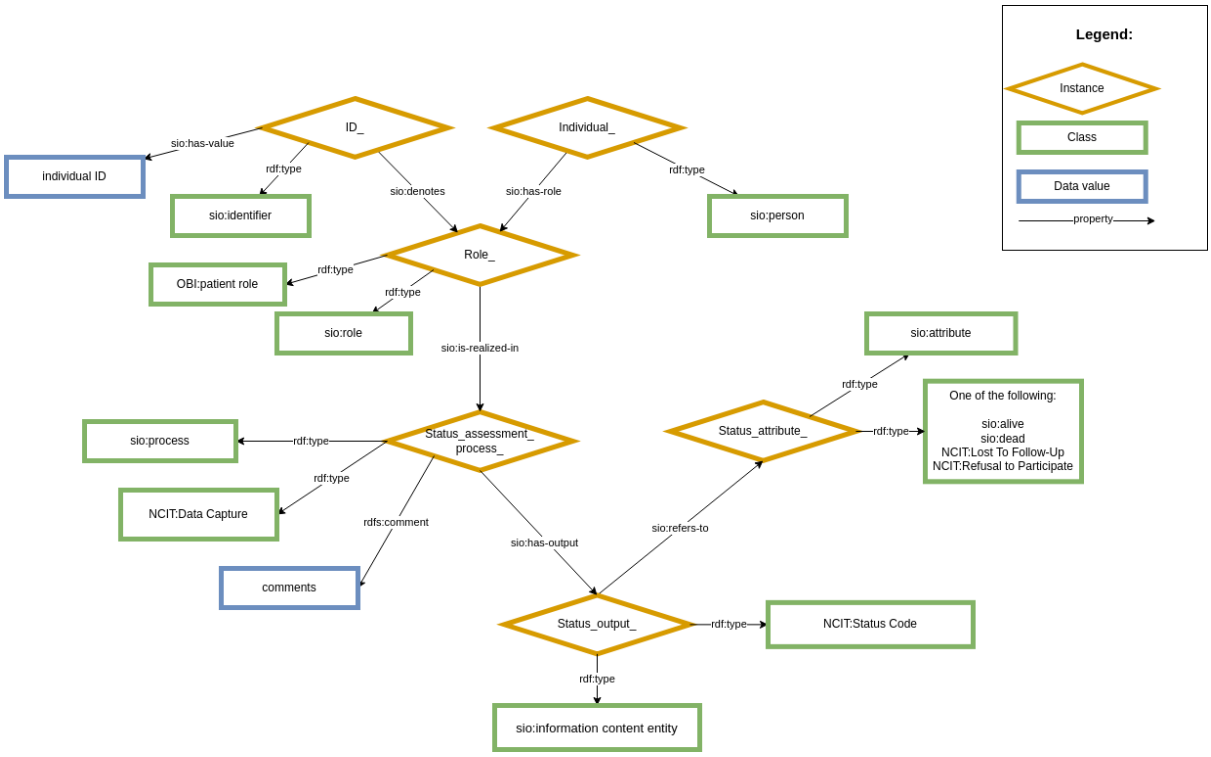


Figure B.4: CARE-SM Participation status information.

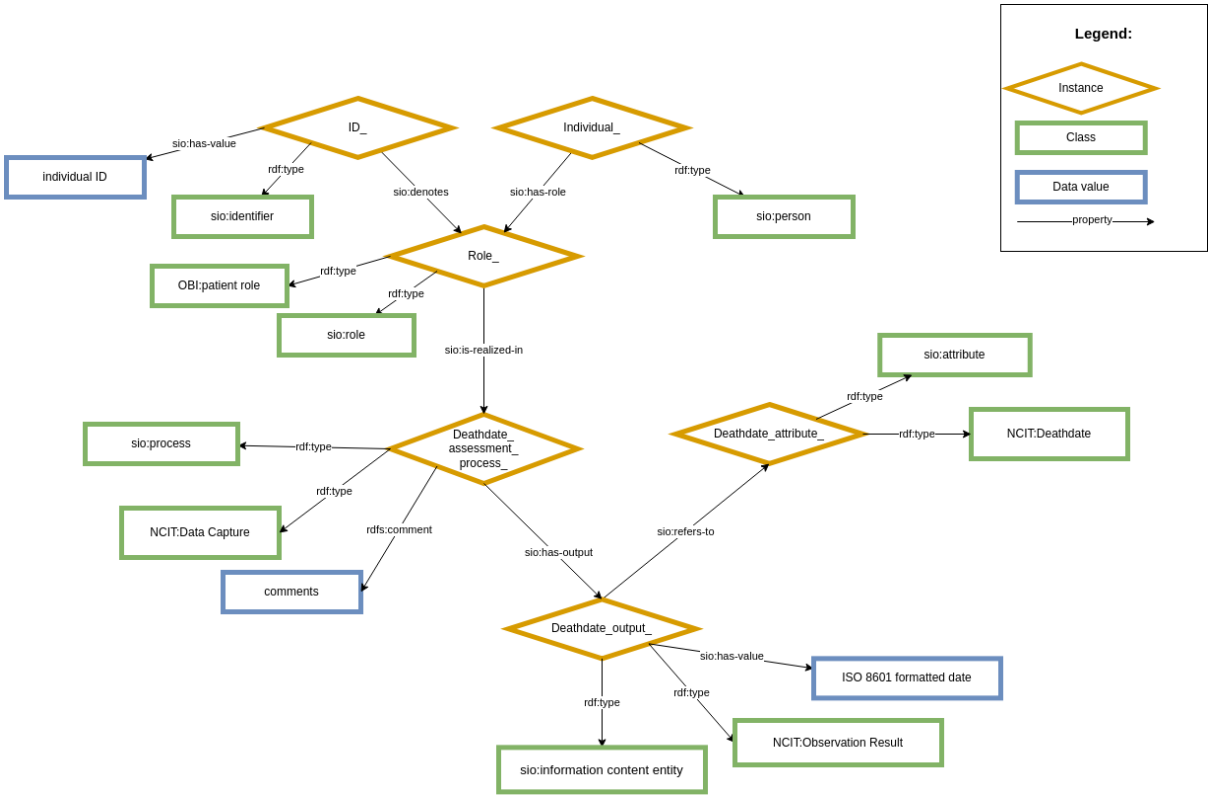


Figure B.5: CARE-SM Deathdate information.

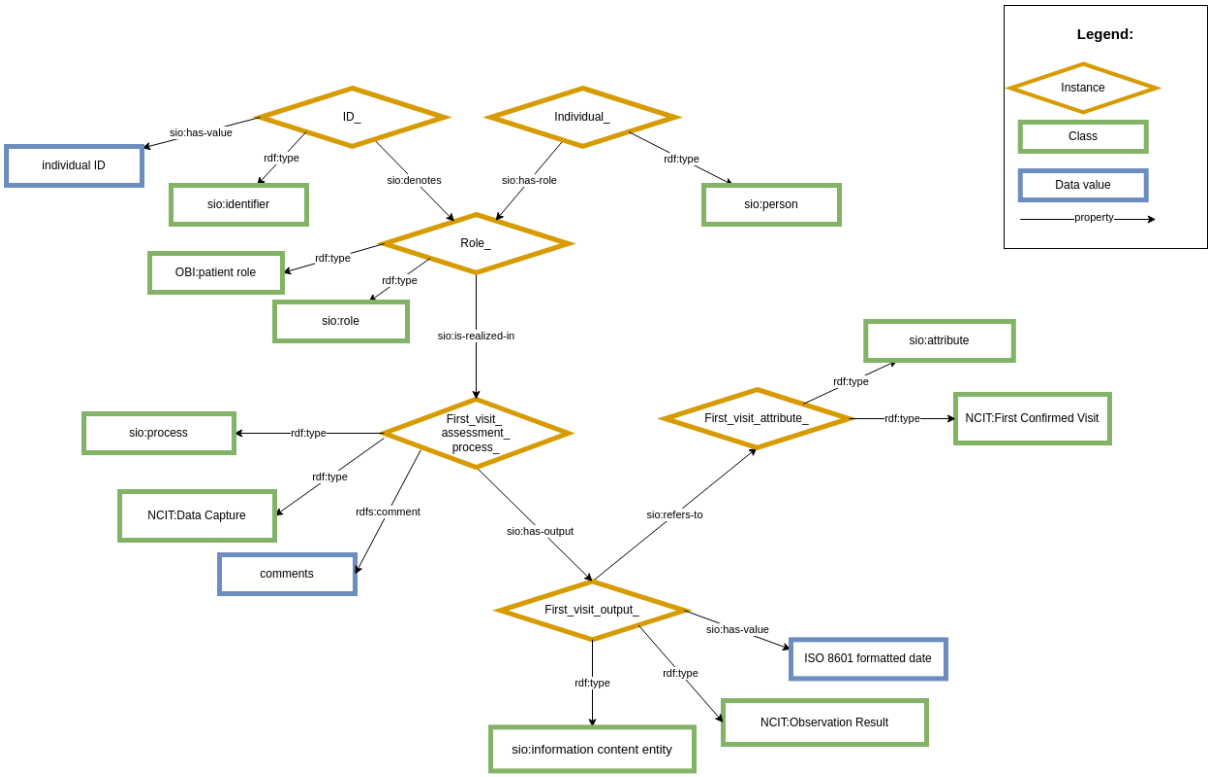


Figure B.6: CARE-SM Care pathway information.

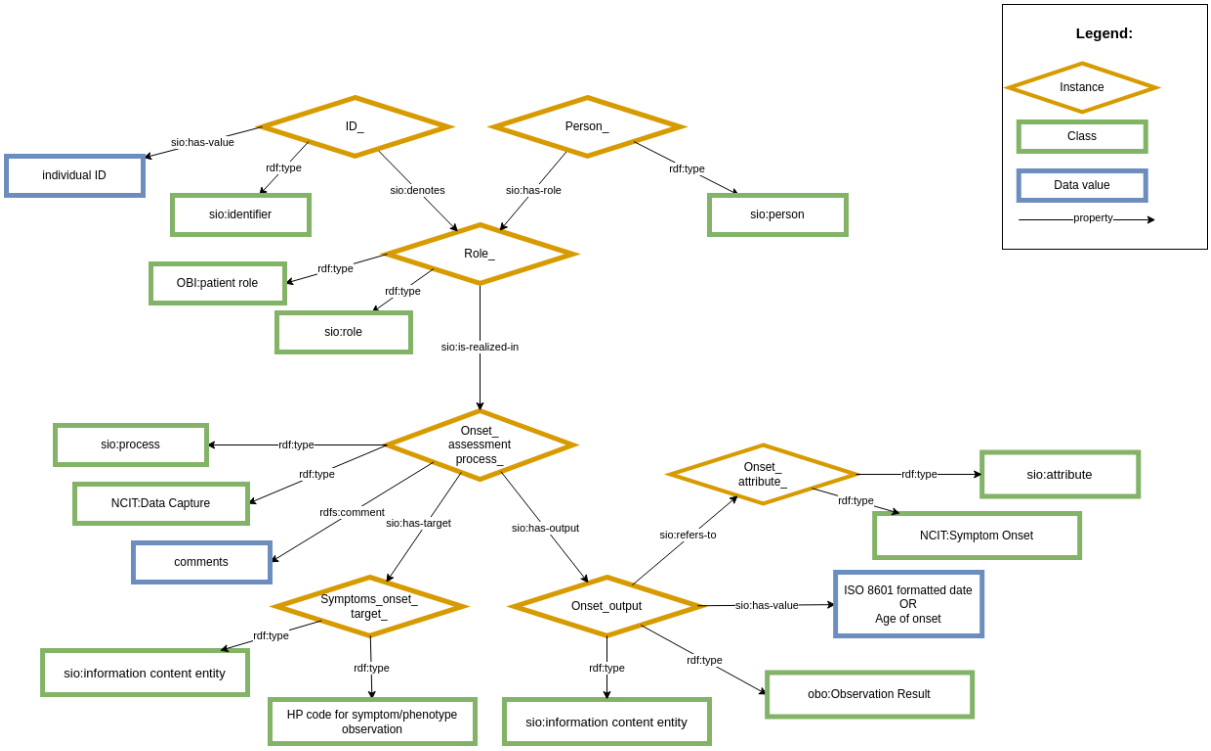


Figure B.7: CARE-SM Symptom onset information.

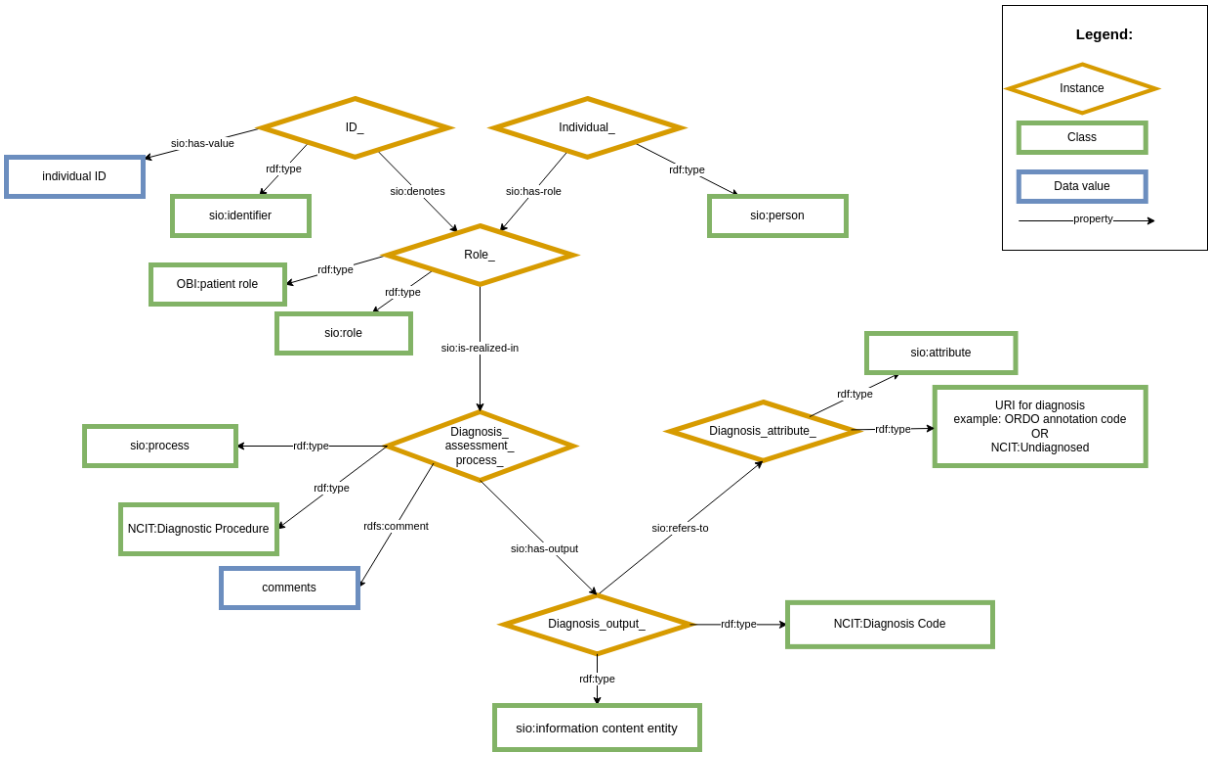


Figure B.8: CARE-SM Diagnosis information.

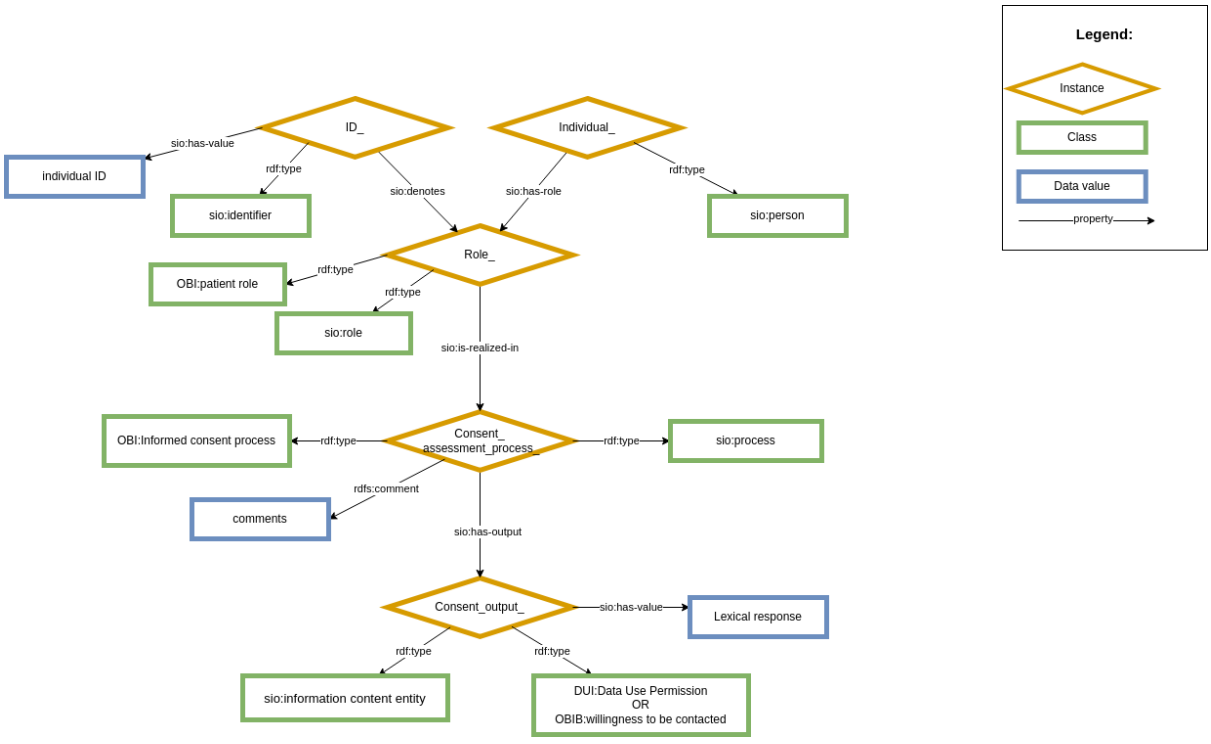


Figure B.9: CARE-SM Consent information.

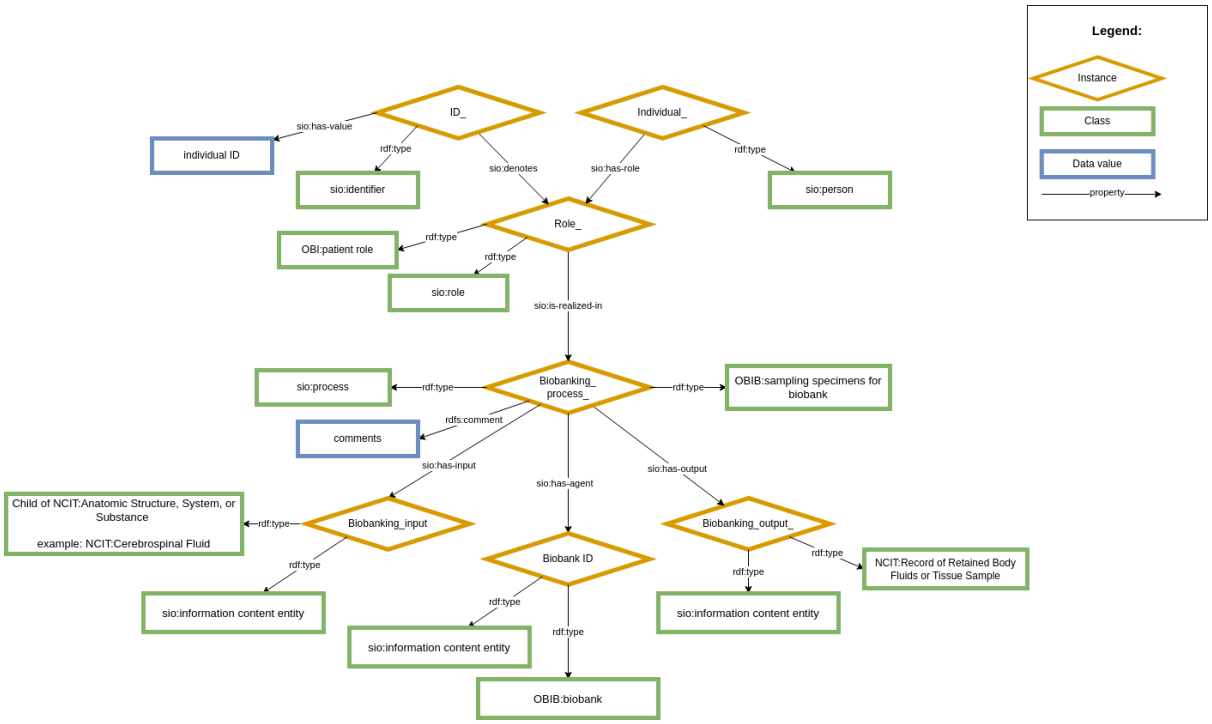


Figure B.10: CARE-SM Biobank information

Annexes C: CARE-SM unified SPARQL query

```
PREFIX obo: <http://purl.obolibrary.org/obo/>
PREFIX sio: <http://semanticscience.org/resource/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX ordo: <http://www.orpha.net/ORDO/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT distinct ?individual_id ?value_startdate ?process_label ?value ?attribute_type
?target_type ?event_id #?activity_type ?substance ?concentration_value ?frequency_value

WHERE {
   GRAPH ?g {
      ?role sio:SIO_000356 ?process ;  a sio:SIO_000016 .
      ?process a sio:SIO_000006, ?process_type; rdfs:label ?process_label; sio:SIO_000229
?output; sio:SIO_000291 ?target; sio:SIO_000230 ?input; sio:SIO_000139 ?agent;
sio:SIO_000339 ?protocol . FILTER(?process_type != sio:SIO_000006) .
      ?output sio:SIO_000628 ?attribute; a sio:SIO_000015; rdfs:label ?output_label.
      ?attribute a sio:SIO_000614.

      OPTIONAL{?output a ?outpute_type. FILTER(?attribute_type != sio:SIO_000015)}
      OPTIONAL{?attribute a ?attribute_type. FILTER(?attribute_type !=
sio:SIO_000614)}
      OPTIONAL{?input a ?input_type. FILTER(?input_type != sio:SIO_000015)}
      OPTIONAL{?target a ?target_type. FILTER(?target_type != sio:SIO_000015)}
```

```
      OPTIONAL{?agent a ?agent_type. FILTER(?agent != sio:SIO_000015)}

      OPTIONAL{?frequency a ?frequency_type, sio:SIO_001367 .FILTER
(?frequency_type != sio:SIO_001367)}
      OPTIONAL{?substance a ?substance_type, sio:SIO_000315 .FILTER
(?substance_type != sio:SIO_000315)}
      OPTIONAL{?concentration a ?concentration_type, sio:SIO_001088 .FILTER
(?concentration_type != sio:SIO_001088)}
      OPTIONAL{?activity a ?activity_type, sio:SIO_000091 .FILTER (?activity_type !=
sio:SIO_000091)}

      OPTIONAL{?process rdfs:comments ?comments}.
      OPTIONAL{?output sio:SIO_000300 ?value}.
      OPTIONAL{?substance sio:SIO_000300 ?substance_value}.
      OPTIONAL{?frequency sio:SIO_000300 ?frequency_value}.
   }
   ?g a obo:NCIT_C62143 ; sio:SIO_000068 ?timeline, ?event ;
                        sio:SIO_000680 ?startdate;
                        sio:SIO_000681 ?enddate;
                        sio:SIO_000687 ?age;
                        sio:SIO_000300 ?uniqid .

   ?age a sio:SIO_001013, obo:NCIT_C25150 .
   ?startdate a sio:SIO_000031, obo:NCIT_C68616 .
   ?enddate a sio:SIO_000032, obo:NCIT_C68617 .
   ?event a obo:NCIT_C25499 ; sio:SIO_000300 ?event_id.

   OPTIONAL{?age sio:SIO_000300 ?age_value .}
   OPTIONAL{?startdate sio:SIO_000300 ?value_startdate .}
   OPTIONAL{?enddate sio:SIO_000300 ?value_enddate .}

   ?timeline a obo:NCIT_C54576, sio:SIO_000417; sio:SIO_000332 ?individual .
   ?individual a sio:SIO_000498, obo:NCIT_C25190 ; sio:SIO_000671
?individual_identifier .
   ?individual_identifier a sio:SIO_000115, obo:NCIT_C164337 ; sio:SIO_000300
?individual_id .
}
```

Annexes D: DQD complete report

STATUS	TABLE	CATEGORY	SUBCATEGORY	LEVEL	NOTES	DESCRIPTION	% RECORDS
FAIL	VISIT_OCCURRENCE	Conformance	Relational	FIELD	None	The number and percent of records that have a value in the VISIT_CONCEPT_ID field in the VISIT_OCCURRENCE table that does not exist in the CONCEPT table. (Threshold=0%).	100.00%
FAIL	CONDITION_ERA	Completeness	None	FIELD	None	The number and percent of records with a value of 0 in the standard concept field CONDITION_CONCEPT_ID in the CONDITION_ERA table. (Threshold=0%).	100.00%
FAIL	CONDITION_OCCURRENCE	Completeness	None	FIELD	None	The number and percent of records with a value of 0 in the standard concept field CONDITION_CONCEPT_ID in the CONDITION_OCCURRENCE table. (Threshold=5%).	100.00%
FAIL	OBSERVATION	Completeness	None	FIELD	None	The number and percent of records with a value of 0 in the standard concept field OBSERVATION_CONCEPT_ID in the OBSERVATION table. (Threshold=5%).	100.00%
FAIL	PROCEDURE_OCCURRENCE	Completeness	None	FIELD	None	The number and percent of records with a value of 0 in the standard concept field PROCEDURE_CONCEPT_ID in the PROCEDURE_OCCURRENCE table. (Threshold=5%).	100.00%

Showing 1 to 5 of 38 entries (filtered from 1,518 total entries) Previous 1 2 3 4 5 ... 8 Next

Figure D1: DQD complete report, part 1/7.

STATUS	TABLE	CATEGORY	SUBCATEGORY	LEVEL	NOTES	DESCRIPTION	% RECORDS
FAIL	DRUG_EXPOSURE	Plausibility	Temporal	FIELD	None	If yes, the number and percent of records with a date value in the DRUG_EXPOSURE_END_DATETIME field of the DRUG_EXPOSURE table that occurs after death. (Threshold=1%).	10.00%
FAIL	DRUG_ERA	Plausibility	Temporal	FIELD	None	The number and percent of records with a date value in the DRUG_ERA_START_DATE field of the DRUG_ERA table that occurs after death. (Threshold=1%).	10.00%
FAIL	DRUG_ERA	Plausibility	Temporal	FIELD	None	The number and percent of records with a date value in the DRUG_ERA_END_DATE field of the DRUG_ERA table that occurs after death. (Threshold=1%).	10.00%
FAIL	DRUG_EXPOSURE	Plausibility	Temporal	FIELD	None	The number and percent of records with a date value in the DRUG_EXPOSURE_START_DATE field of the DRUG_EXPOSURE table that occurs after death. (Threshold=1%).	10.00%
FAIL	DRUG_EXPOSURE	Plausibility	Temporal	FIELD	None	The number and percent of records with a date value in the DRUG_EXPOSURE_START_DATETIME field of the DRUG_EXPOSURE table that occurs after death. (Threshold=1%).	10.00%

Showing 11 to 15 of 38 entries (filtered from 1,518 total entries) Previous 1 2 3 4 5 ... 8 Next

Figure D2: DQD complete report, part 2/7.

STATUS	TABLE	CATEGORY	SUBCATEGORY	LEVEL	NOTES	DESCRIPTION	% RECORDS
FAIL	DRUG_EXPOSURE	Plausibility	Temporal	FIELD	None	The number and percent of records with a date value in the DRUG_EXPOSURE_END_DATE field of the DRUG_EXPOSURE table that occurs after death. (Threshold=1%).	10.00%
FAIL	DRUG_EXPOSURE	Plausibility	Temporal	FIELD	None	The number and percent of records with a date value in the DRUG_EXPOSURE_END_DATETIME field of the DRUG_EXPOSURE table that occurs after death. (Threshold=1%).	10.00%
FAIL	CONDITION_OCCURRENCE	Plausibility	Temporal	FIELD	None	If yes, the number and percent of records with a date value in the CONDITION_START_DATE field of the CONDITION_OCCURRENCE table that occurs after death. (Threshold=1%).	4.35%
FAIL	CONDITION_OCCURRENCE	Plausibility	Temporal	FIELD	None	If yes, the number and percent of records with a date value in the CONDITION_START_DATETIME field of the CONDITION_OCCURRENCE table that occurs after death. (Threshold=1%).	4.35%
FAIL	CONDITION_OCCURRENCE	Plausibility	Temporal	FIELD	None	If yes, the number and percent of records with a date value in the CONDITION_END_DATE field of the CONDITION_OCCURRENCE table that occurs after death. (Threshold=1%).	4.35%

Showing 16 to 20 of 38 entries (filtered from 1,518 total entries) Previous 1 2 3 4 5 ... 8 Next

Figure D3: DQD complete report, part 3/7.

STATUS	TABLE	CATEGORY	SUBCATEGORY	LEVEL	NOTES	DESCRIPTION	% RECORDS
FAIL	CONDITION_OCCURRENCE	Plausibility	Temporal	FIELD	None	If yes, the number and percent of records with a date value in the CONDITION_END_DATETIME field of the CONDITION_OCCURRENCE table that occurs after death. (Threshold=1%).	4.35%
FAIL	CONDITION_OCCURRENCE	Plausibility	Temporal	FIELD	None	The number and percent of records with a date value in the CONDITION_START_DATE field of the CONDITION_OCCURRENCE table that occurs after death. (Threshold=1%).	4.35%
FAIL	CONDITION_OCCURRENCE	Plausibility	Temporal	FIELD	None	The number and percent of records with a date value in the CONDITION_START_DATETIME field of the CONDITION_OCCURRENCE table that occurs after death. (Threshold=1%).	4.35%
FAIL	CONDITION_OCCURRENCE	Plausibility	Temporal	FIELD	None	The number and percent of records with a date value in the CONDITION_END_DATE field of the CONDITION_OCCURRENCE table that occurs after death. (Threshold=1%).	4.35%
FAIL	CONDITION_OCCURRENCE	Plausibility	Temporal	FIELD	None	The number and percent of records with a date value in the CONDITION_END_DATETIME field of the CONDITION_OCCURRENCE table that occurs after death. (Threshold=1%).	4.35%

Showing 21 to 25 of 38 entries (filtered from 1,518 total entries) Previous 1 ... 4 5 6 7 8 Next

Figure D4: DQD complete report, part 4/7.

STATUS	TABLE	CATEGORY	SUBCATEGORY	LEVEL	NOTES	DESCRIPTION	% RECORDS
FAIL	CONDITION_ERA	Plausibility	Temporal	FIELD	None	If yes, the number and percent of records with a date value in the CONDITION_ERA_START_DATE field of the CONDITION_ERA table that occurs after death. (Threshold=1%).	3.51%
FAIL	CONDITION_ERA	Plausibility	Temporal	FIELD	None	The number and percent of records with a date value in the CONDITION_ERA_START_DATE field of the CONDITION_ERA table that occurs after death. (Threshold=1%).	3.51%
FAIL	CONDITION_ERA	Plausibility	Temporal	FIELD	None	The number and percent of records with a date value in the CONDITION_ERA_END_DATE field of the CONDITION_ERA table that occurs after death. (Threshold=1%).	3.51%
FAIL	VISIT_OCCURRENCE	Plausibility	Temporal	FIELD	None	If yes, the number and percent of records with a date value in the VISIT_START_DATE field of the VISIT_OCCURRENCE table that occurs after death. (Threshold=1%).	2.21%
FAIL	VISIT_OCCURRENCE	Plausibility	Temporal	FIELD	None	If yes, the number and percent of records with a date value in the VISIT_START_DATETIME field of the VISIT_OCCURRENCE table that occurs after death. (Threshold=1%).	2.21%

Showing 26 to 30 of 38 entries (filtered from 1,518 total entries) Previous 1 ... 4 5 6 7 8 Next

Figure D5: DQD complete report, part 5/7.

STATUS	TABLE	CATEGORY	SUBCATEGORY	LEVEL	NOTES	DESCRIPTION	% RECORDS
FAIL	VISIT_OCCURRENCE	Plausibility	Temporal	FIELD	None	If yes, the number and percent of records with a date value in the VISIT_END_DATE field of the VISIT_OCCURRENCE table that occurs after death. (Threshold=1%).	2.21%
FAIL	VISIT_OCCURRENCE	Plausibility	Temporal	FIELD	None	If yes, the number and percent of records with a date value in the VISIT_END_DATETIME field of the VISIT_OCCURRENCE table that occurs after death. (Threshold=1%).	2.21%
FAIL	VISIT_OCCURRENCE	Plausibility	Temporal	FIELD	None	The number and percent of records with a date value in the VISIT_START_DATE field of the VISIT_OCCURRENCE table that occurs after death. (Threshold=1%).	2.21%
FAIL	VISIT_OCCURRENCE	Plausibility	Temporal	FIELD	None	The number and percent of records with a date value in the VISIT_START_DATETIME field of the VISIT_OCCURRENCE table that occurs after death. (Threshold=1%).	2.21%
FAIL	VISIT_OCCURRENCE	Plausibility	Temporal	FIELD	None	The number and percent of records with a date value in the VISIT_END_DATE field of the VISIT_OCCURRENCE table that occurs after death. (Threshold=1%).	2.21%

Showing 31 to 35 of 38 entries (filtered from 1,518 total entries) Previous 1 ... 4 5 6 7 8 Next

Figure D6: DQD complete report, part 6/7.

STATUS	TABLE	CATEGORY	SUBCATEGORY	LEVEL	NOTES	DESCRIPTION	% RECORDS
FAIL	VISIT_OCCURRENCE	Plausibility	Temporal	FIELD	None	The number and percent of records with a date value in the VISIT_END_DATETIME field of the VISIT_OCCURRENCE table that occurs after death. (Threshold=1%).	2.21%
FAIL	MEASUREMENT	Plausibility	Temporal	FIELD	None	The number and percent of records with a date value in the MEASUREMENT_DATE field of the MEASUREMENT table that occurs after death. (Threshold=1%).	1.75%
FAIL	MEASUREMENT	Plausibility	Temporal	FIELD	None	The number and percent of records with a date value in the MEASUREMENT_DATETIME field of the MEASUREMENT table that occurs after death. (Threshold=1%).	1.75%

Showing 36 to 38 of 38 entries (filtered from 1,518 total entries) Previous 1 ... 4 5 6 7 8 Next

Figure D7: DQD complete report, part 7/7.

Annexes E: Scientific collaborations and first-author papers during my thesis.

E.1: Kaliyaperumal et al, 2022

Rajaram Kaliyaperumal; Mark D. Wilkinson; Pablo Alarcón Moreno; Nirupama Benis; Ronald Cornet; Bruna dos Santos Vieira; Michel Dumontier; César Henrique Bernabé; Annika Jacobsen; Clémence M. A. Le Cornec et al. (2023). Semantic modelling of Common Data Elements for Rare Disease registries, and a prototype workflow for their deployment over registry data. Journal of Biomedical Semantics.

E.2: Alarcon et al., 2023

Alarcon, P., Braun, I., Hartley, E., Olson, D., Benis, N., Cornet, R., Wilkinson, M. & Walls, R. L., (2023) "Leveraging Biolink as a "Rosetta Stone" Between C-Path and EJP-RD Semantic Models Provides Emergent Interoperability", Journal of the Society for Clinical Data Management 3(1).

E.3: Damme et al., 2023

Philip van Damme; Pablo Alarcón Moreno; César H. Bernabé; Alberto Cámara Ballesteros; Clémence M. A. Le Cornec; Bruna Dos Santos Vieira; K. Joeri van der Velde; Shuxin Zhang; Claudio Carta; Ronald Cornet et al. (2023) A Resource for Guiding Data Stewards to Make European Rare Disease Patient Registries FAIR. Data Science Journal.

Journal of Biomedical Semantics

RESEARCH Open Access

Semantic modelling of common data elements for rare disease registries, and a prototype workflow for their deployment over registry data

Rajaram Kaliyaperumal[1†], Mark D. Wilkinson[2*†], Pablo Alarcón Moreno[2], Nirupama Benis[3], Ronald Cornet[3], Bruna dos Santos Vieira[5,4], Michel Dumontier[6], César Henrique Bernabé[1], Annika Jacobsen[1], Clémence M. A. Le Cornec[7], Mario Prieto Godoy[2], Núria Queralt-Rosinach[1], Leo J. Schultze Kool[4], Morris A. Swertz[8], Philip van Damme[3], K. Joeri van der Velde[8], Nawel Lalout[5,9], Shuxin Zhang[3] and Marco Roos[1]

Abstract

Background: The European Platform on Rare Disease Registration (EU RD Platform) aims to address the fragmentation of European rare disease (RD) patient data, scattered among hundreds of independent and non-coordinating registries, by establishing standards for integration and interoperability. The first practical output of this effort was a set of 16 Common Data Elements (CDEs) that should be implemented by all RD registries. Interoperability, however, requires decisions beyond data elements - including data models, formats, and semantics. Within the European Joint Programme on Rare Diseases (EJP RD), we aim to further the goals of the EU RD Platform by generating reusable RD semantic model templates that follow the FAIR Data Principles.

Results: Through a team-based iterative approach, we created semantically grounded models to represent each of the CDEs, using the SemanticScience Integrated Ontology as the core framework for representing the entities and their relationships. Within that framework, we mapped the concepts represented in the CDEs, and their possible values, into domain ontologies such as the Orphanet Rare Disease Ontology, Human Phenotype Ontology and National Cancer Institute Thesaurus. Finally, we created an exemplar, reusable ETL pipeline that we will be deploying over these non-coordinating data repositories to assist them in creating model-compliant FAIR data without requiring site-specific coding nor expertise in Linked Data or FAIR.

* Correspondence: mark.wilkinson@upm.es
†Rajaram Kaliyaperumal and Mark D. Wilkinson contributed equally to this work.
[2]Departamento de Biotecnología-Biología Vegetal, Escuela Técnica Superior de Ingeniería Agronómica, Alimentaria y de Biosistemas, Centro de Biotecnología y Genómica de Plantas (CBGP), Universidad Politécnica de Madrid (UPM), Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA), Pozuelo de Alarcón, Madrid, ES, Spain
Full list of author information is available at the end of the article

Conclusions: Within the EJP RD project, we determined that creating reusable, expert-designed templates reduced or eliminated the requirement for our participating biomedical domain experts and rare disease data hosts to understand OWL semantics. This enabled them to publish highly expressive FAIR data using tools and approaches that were already familiar to them.

Keywords: FAIR data, Rare disease, Interoperability, Linked data, Data transformation, Semantic web, Ontologies, Common data elements, Disease registries

Background

The FAIR Principles [1] aim to provide guidance that will lead to an internet of data and services that is highly descriptive and machine-accessible, resulting in more extensive data discovery and reuse. FAIR (Findable, Accessible, Interoperable, and Reusable) data requires unambiguously identified entities to be richly described by unambiguously defined and identified concepts from thesauri and ontologies that are widely shared within a community and machine readable. When this is achieved, it will become much more straightforward to discover task-relevant data over distributed sites, accurately integrate those data, or analyse them by 'data visiting'.

A significant barrier to Rare Disease (RD) research is that RD data is (a) extremely scarce, and (b) spread over many "boutique" repositories, often single-disease-specific and often curated by biomedically-oriented experts, who may not have access to experts in data or knowledge representation, capture or archival. In an initial step to address this, the EU RD Platform has begun to establish standards for integration and interoperability. The first practical output of this effort was a set of 16 Common Data Elements that should be implemented by all RD registries [2]. These include facets such as "sex", "date of birth", "age of onset", and "diagnosis", often together with a constraint on the allowed values of each of these data elements (for example, the possible values of 'age at onset' are 'Antenatal', 'At birth', 'Date (dd/mm/yyyy)', or 'Undetermined'). Achieving uniformity of these 16 data facets, over all RD registries and biobanks, would be an excellent first-step towards enhanced discovery and reuse of these precious data. Web-scale – which implies "mechanized" – interoperability, however, requires decisions beyond just a list of data elements, including data models, formats, and semantics.

The European Joint Programme on Rare Diseases (EJP RD) is an expansive European (with foreign partners) project aiming to reduce the suffering of rare disease patients and their families, through technical, clinical, social, economic, and health-services mechanisms. EJP RD spans 35 countries with 87 beneficiaries and 52 linked parties, and spans all 24 European Rare Disease Reference Networks (ERNs), totalling approximately 1200 people. Each ERN focuses on a particular class of rare diseases (for example, neuromuscular or vascular), and thus each ERN will have multiple registries and/or biobanks, which will vary in their level of systematic and schematic coordination even within a single ERN. To improve the utility of this massive ecosystem of data collection and curation, EJP RD aims to further the goals of the EU RD Platform by generating a "Virtual Platform" for interoperability between RD data assets throughout Europe and beyond. Speaking only of the technology layer, the Virtual Platform will provide common, harmonized access to discovery of task-relevant data resources, supported by a rich layer of metadata describing the content and context of each participating ERN repository. In part, this is being pursued by generating metadata that follows the FAIR Data Principles [3, 4], and global metadata standards are well-established (e.g., Dublin Core [5] and Data Catalog – DCAT [6]). This cannot be said for data, however. The diversity of data, and the wide range of mechanisms, tools, and devices for generating it often thwart generic approaches to creation of data schema. Nevertheless, historically, within the RD community, there have been efforts to train individual data custodians to create FAIR data at-source. These have taken the form of annually recurring "Bring Your Own Data" [7] workshops (BYODs) where data custodians meet FAIR experts and get hands-on experience in making their resources FAIR.

Because of their open-ended, exploratory structure, these BYOD events did not converge on a unified model for RD data, nor even the elements that should be included in those models. As such, the workshops primarily succeeded in raising awareness of FAIR, and the utility and benefits of following the FAIR Principles; however, the degree of inter-repository harmonization achieved by these workshops was extremely limited. Nevertheless, some preliminary data models [8] were created at BYOD workshops, including the early version (V0.1.0) of CDE semantic model that is the focus of this manuscript, which was developed during the FAIRification of a registry for vascular anomalies [9–11]..

In the case of the EJP RD project, it was immediately clear that training individual participants in FAIR data modelling themselves would be challenging for many reasons - RD registries are limited by funding, FAIR

expertise, and time. All three of those barriers make it infeasible for the initial FAIRification pathway for EJP RD to involve significant decision-making by the resource custodian. Rather, we decided to centralize many of the decisions, ensuring that they were made by a small group of FAIR experts, and then disseminated outward to the individual participating registries and biobanks via a layer of registry "liaisons" who would communicate the needs, in both directions, between the data modellers and the registry custodians.

The final problem was how to enact the FAIRification itself - that is, how to do the "extract" and "transform" portions of the traditional Extract/Transform/Load (ETL) pipeline over resources that had no coordinating structure, and potentially no ability to code data transformation software themselves. Thus, we needed to identify an ETL pipeline that could be deployed anywhere, over any native data structure, in highly secure privacy-sensitive environments, and execute a successful transformation using only the expertise that could be expected of most repository curators.

Here we describe the process of data modelling within the EJP RD, as applied to the set of CDEs defined by the EU RD Platform. We describe the semantic basis of those models, and how they have already been applied to distinctly different data types, showing that they have not been overly "fitted" to the data elements defined by the CDEs. Finally, we describe our current attempts to build an ETL pipeline that can fill these models, using a simple, structured Comma-Separated Value (CSV) export of source data from the originating registry hosts. To help orient a broad range of readers, we attempt to split the discussion into three groupings: "FAIR Expert Activities", where technical details and decisions are described; "Data Custodian Activities", where less technical deployment decisions and activities are discussed; and where relevant, "Data Steward Activities" where the role of the "liaisons" between the FAIR experts and the data custodians are highlighted.

Methods

FAIR expert activities - modelling

Modelling activities were undertaken via weekly meetings of a core group of EJP RD researchers with extensive experience in ontologies, knowledge representation, Linked Data modelling, and FAIR data. Meetings were carried out via Microsoft Teams, where the model under discussion was presented via screen sharing.

As noted above, the European Platform on Rare Disease Registration has determined a set of 16 CDEs for RD registration. These are detailed in Table 1.

Using these elements as a guide, together with additional documentation detailing how these elements should be filled, a first pass modelling phase [9] was undertaken where Linked Data representations for each CDE were constructed, using existing ontological terms

Table 1 The European Platform for Rare Disease Registration set of Common Data Elements that should be made available by all rare disease registries

Element ID	Name	Values
1.1	Pseudonym	String
2.1	Date of birth	dd/mm/yyyy
2.2	Sex	Female, Male, Undetermined, Foetus (Unknown)
3.1	Patient Status	Alive, Dead, Lost in Follow-up, Opted-out
3.2	Date of Death	dd/mm/yyyy
4.1	First contact with specialized centre	dd/mm/yyyy
5.1	Age at onset	Antenatal, At Birth, Date (dd/mm/yyyy), Undetermined.
5.2	Age at diagnosis	Antenatal, At Birth, Date, Undetermined
6.1	Diagnosis of the rare disease	ORPHA Code, Alpha Code, ICD9/10 Code, ICD9-CM Code
6.2	Genetic Diagnosis	Human Genome Variant Sequence (HGVS), HUGO Gene Nomenclature Committee (HGNC), Online Medelian Inheritance in Man (OMIIM) Codes
6.3	Undiagnosed case	Human Phenotype Ontology code and/or HGVS Code related to the inability to diagnose.
7.1	Agreement to be contacted for research purposes	Yes/No
7.2	Consent to reuse data	Yes/No
7.3	Biological Sample?	Yes/No
7.4	Biobank?	URL/No
8.1	Disability Classification via International Classification of Functioning and Disability (ICF)	Score

or other shared Globally Unique Identifiers (GUID) wherever possible to model, for example, genotypes (OMIM Codes [12]) and phenotypes (Human Phenotype Ontology Codes [13]).

These first-pass models were then used to frame a more conceptual modelling process, looking at (for example) the inter-dependencies between the CDEs, the "nature" of the data – for example, is it obtained by questionnaire or by physical examination? – and what additional annotation would be useful to contextualize the CDE for correct interpretation (e.g., the dates of various phenotype onsets could be used to build a longitudinal record of the patient's response to treatment). There were several over-arching guidelines that constrained this modelling process:

1) We should use the minimum number of ontologies possible.

2) We must strictly adhere to the ontological definition of a concept.

3) The ontologies/vocabularies used must not have a restrictive license.

4) The model should be designed in a forward-looking manner, anticipating other likely data elements, to minimize the need for future disruptive changes.

Examination of the EU RD CDEs revealed that there were, in fact, many inter-dependencies between them – meaning that one CDE could not reliably be understood or contextualized without one or more of the others. For example, CDE 1.1 - the patient pseudonym – must be a part of every CDE, since all CDEs are related to an individual patient. For example, 4.1 'First contact with a specialized centre' cannot be interpreted without a reference to the patient that made contact with the centre (via their pseudonym). Similarly, since individuals may have multiple diseases, each with its own diagnosis (CDE 6.1), the "age at diagnosis" (CDE 5.2) must somehow relate to the disease which was diagnosed at that age. In addition, we noted that most CDEs focused on data that would result from a formal interaction in a clinical setting, but those data gathering processes might be undertaken at different locations and times. For example, obtaining a biological specimen (CDE 7.3) would often be a surgical process, which would be undertaken in entirely different circumstances than the administration of a questionnaire to generate a disability score (CDE 8.1). Although the CDE requirements from the EU RD Platform do not require that this metadata be represented, it is nevertheless true that these details likely are being captured in many cases, thus it is useful to plan a model that can carry this contextual metadata, now or in the future. This minimizes the degree to which EJP RD participants would have to change their workflows to adapt to future changes.

We examined two ontologies that are commonly used in the life sciences to model processes, workflows, and their participants. The Provenance ontology (PROV-O) [14] aims to capture information about the sequence of events that leads to an output, such as who executed which version of which algorithm at what time, using what input data, and where is the output data stored. The Semantic Science Integrated Ontology (SIO) [15] is also capable of modelling entities, processes, and their qualities/attributes, but includes additional entity-to-entity or entity-to-process relationships that enable rigorous and highly explicit machine-readable patterns associating these to one another. SIO also has domain extensions, including biology and bioinformatics, that can help ensure that many clinical or biological concepts are being used in a logically sound manner. For example, SIO includes CDE-relevant concepts such as "medical diagnosis", which allows us to use SIO-defined properties and entities for the majority of the CDEs. Finally, SIO has the capacity to represent data content – that is, while PROV-O has the concept of an Entity, which could represent the output of a process, it does not have the ability to describe what that entity is, or its value, or its measurement units. Finally, PROV-O has no way of representing the attributes or qualities of an individual. Given that almost all of the CDEs are measurements of some attribute of the patient who participated in a clinical process, the inability to associate the output of a process (like a phenotype) as an attribute of the participating patient would make a PROV-O model highly unintuitive for our target end-users. As such, while PROV-O might be useful at a later date to describe, for example, the precise details of a Phenotyping or Genetic Diagnosis workflow, our needs in this modelling exercise are distinct, and are better represented by SIO concepts.

Following the documented design patterns for SIO [16] we derived the core model shown in Fig. 1. Some of the rationale for this model are as follows: All CDE observations are, in some way "about" an individual patient. As such, it is necessary to connect patients to these observations. In some cases, the CDEs pertain to a direct attribute of the patient (e.g., their birth date). In other cases, the CDE is not an attribute of the patient per se, but rather the connection between a patient and the CDE is via an action or activity that the patient engaged in; for example, the first interaction of a patient with a rare disease expert centre. Certainly, for all CDEs there is at least the process of recording the information, and as such, we decided that a "process" was a concept shared by all CDEs. Early discussions also raised the issue of an individual having multiple roles in the healthcare system, for example, being both a patient and a

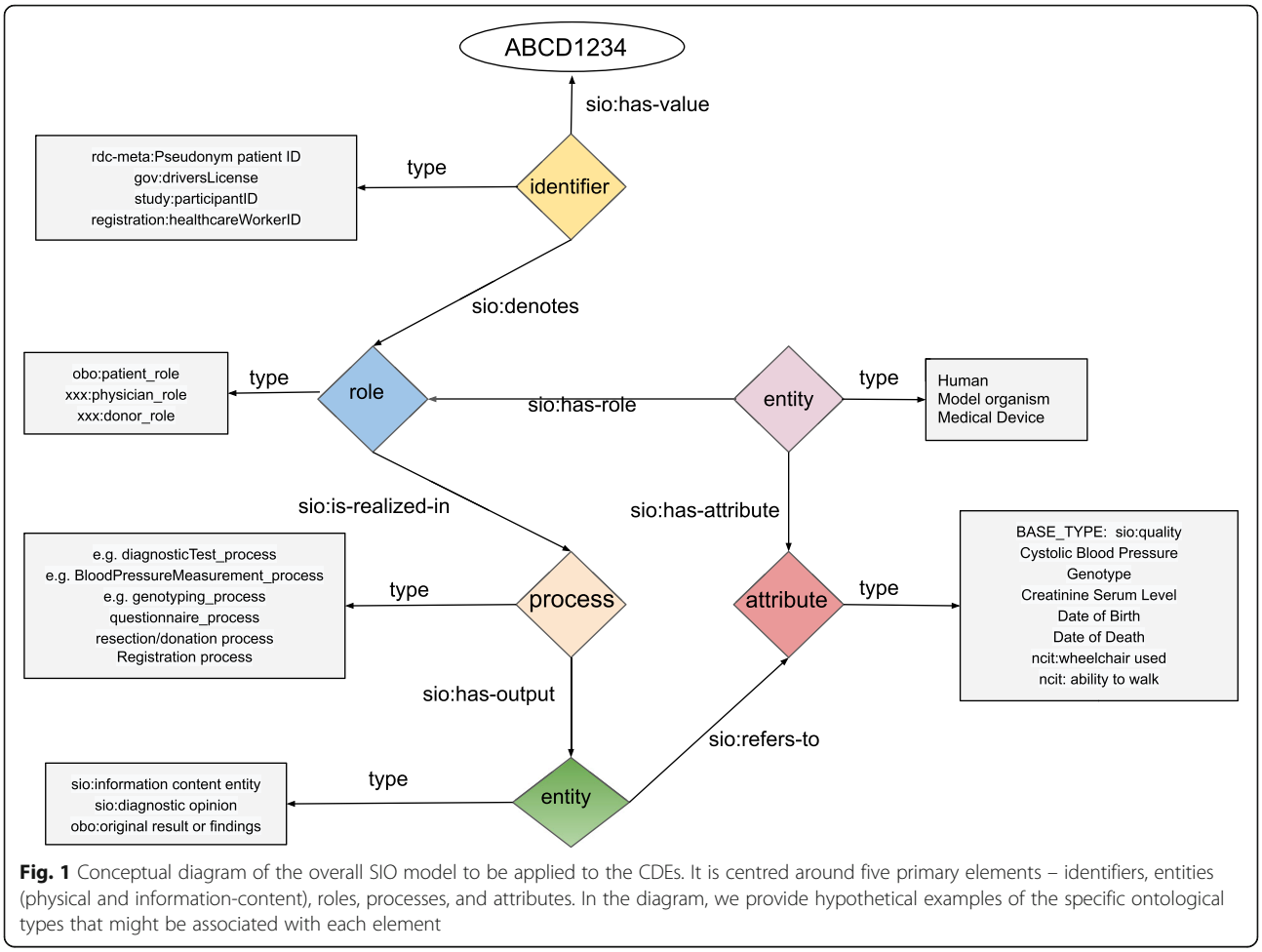


Fig. 1 Conceptual diagram of the overall SIO model to be applied to the CDEs. It is centred around five primary elements – identifiers, entities (physical and information-content), roles, processes, and attributes. In the diagram, we provide hypothetical examples of the specific ontological types that might be associated with each element

physician. As such, it was necessary to connect a participant to the process indirectly, by declaring the role they play in the process. An individual may have many roles, and we determined that in every case, there was a distinct identifier that was assigned to that role – for example, a driver's license ID is assigned to one's role as a driver, and a student ID is assigned to one's role as a student, yet both identifiers may apply to the same individual. As such, we associated identifiers with individuals via their role, rather than directly. Finally, processes have outputs, where those outputs (often) refer to some measurement of an attribute of the patient. The attribute, and its measurement, are distinct – for example, all patients share the attribute of "sex" but for some patients this attribute has the value "male" and for others it has the value "female".

Combining these considerations leads to the core model shown in Fig. 1, where there are 5 "kinds" of things: entities (individuals, and measurements), roles, processes, attributes, and identifiers. While there are additional relationships between these concepts, we removed all but the relations required to connect the model. This will simplify the creation of query systems, by limiting the possible ways the model can be explored, better enabling the construction of reusable query templates (an activity that is also being undertaken within the Virtual Platform, to ensure that end-users do not need to learn a formal query language for these data).

Using this high-level model as a guide, the EJP RD semantic modelling group then reiterated the process of examining each CDE and, through Teams meetings and dedicated "designathons" we reached agreement on which portions of the high-level model were appropriate for each CDE, and what the ontological type constraints (square boxes in Fig. 1) should be for the elements of that specific CDE model. As part of the modelling process, we selected a "base type" for each of the model elements, for example, the process node is always ontologically typed as a "sio:process". In this way, if there is not a more specific type assigned to the model node, we still maintain the best practice of having all nodes in our model ontologically typed.

These "base types" are built into our transformation templates (described below) and require no knowledge by the end-user. Finally, where appropriate, we selected ontological concepts that would be allowed as values or attributes for various CDEs. For example, in the personal information CDE we selected the National Cancer Institute Thesaurus' [17] terms for "male" and "female", and we selected the Human Phenotype Ontology [13] as the set of possible values for the phenotypic diagnosis CDE.

FAIR expert activities – design a "lingua franca" for data extraction

Registries participating in the EJP RD have a wide range of underlying infrastructures and data management and curation expertise, ranging from well-established commercial enterprises such as Castor [18], to smaller, parent-run organizations even using spreadsheets to capture data. They are also largely not coordinating with one another and are therefore making independent decisions about database structures and the formats of the captured data. It was therefore clear that, as an initial step to harmonization, we needed to find a "lowest common denominator" for an intermediate data representation format – something more predictable than the various source data structures, but not yet FAIR. It needed to be a format that can be generated by data custodians at any level of expertise, from any starting format, hopefully using only tools with which they are already familiar. Moreover, a primary objective of EJP RD is to encourage FAIRness beyond the EJP RD itself, thus the selected format should be applicable to a wide variety of expert domains and situations.

Through discussions with EJP RD partners, it became clear that there was a preference for very straightforward data structures such as CSV, since this is an exchange format that can be derived easily from any of the more complex formats. As such, CSV was selected as the "lingua franca" that would be used by all participants as an export format, as a step towards harmonization and FAIRness. The rules given to the data custodians explaining how to generate these CSV files is described in the section "**Data Custodian Activities – Generating template-compliant CSV**".

FAIR expert activities – CSV to RDF mapping

Having selected CSV as a starting format, we then chose a mapping framework that could transform CSV data into RDF. RDF Mapping Language (RML) was the selected technology, as it is capable of modelling reusable templates that support not only CSV to RDF transformations, but also transformations from other formats, allowing us the opportunity to increase in complexity in the future without dramatically changing our pipeline.

RML templates specify individual triple patterns that should be created during a transformation. The subject Uniform Resource Identifier (URI), predicate URI, and object URI are represented as strings that may contain variables, where the variables are references to locations within the source document (e.g., the appropriate column header within a CSV file). During a transformation, every variable in an RML template is replaced by the value of that location within a single source record (e.g., a single row of a CSV file) and then the source is iterated over all records to complete the transformation. RML templates themselves are represented in RDF and are therefore not always easily human-readable. With the aim of simplifying the RML syntax, such that our EJP RD FAIRification stewards, or potentially the registry data custodians themselves, could edit the template if required, we identified a second, related technology – YARRRML [19] – which is a more human-readable way to declare RML transformation rules, using YAML as the syntax. YARRRML documents can be converted into RML templates, which can then be automatically applied to CSV files to achieve their transformation. Having selected these tools, the FAIRification experts then transcribed each of the CDE models into a set of YARRRML rules that were then executed to generate RML mapping documents. The YARRRML documents are stored in the "YARRRML_Transform_Templates" folder of the CDE Project GitHub [20] for others to explore and reuse. The final step of modelling was to create documentation and example data to provide to the registry custodians, to guide them in the requirements for the CSV files.

Data custodian activities – generating template-compliant CSV

The process for creation of the CSV files for each CDE will likely differ for each registry, as their individual situations will be diverse. Based on the documentation and examples provided by the FAIR Expert team, considerations for the registry custodians include:

- Ensuring date formatting is correct (ISO 8601)
- Ensuring that any abbreviated ontology terms have been converted into their equivalent full URIs (e.g., Human Phenotype Ontology terms must be represented by their URI in the CSV file)
- Ensuring any data elements have been modified to match the documented constraints (e.g., conversion of textual descriptors into ontology term URIs)
- Ensuring that every row is uniquely identified (for this purpose, we have established a web service that can be called from MS Excel or a custom script that generates a unique identifier based on a timestamp, since MS Excel has no inherent capability to generate GUIDs without custom coding)

Data steward activities – assisting data custodians

In anticipation of the data custodians having questions about how to generate template-compliant CSV and the rationale behind certain decisions, every participant was assigned a FAIR Data Steward to provide them with assistance and advice. For example, we could anticipate data custodians being concerned about why one ontology was chosen over another or needing advice on a tool that can cleanse or edit CSV files. In addition, the FAIR Stewards would bring back suggestions from the Custodians to the FAIR Expert meetings, providing a useful feedback mechanism where the Steward had personally engaged with the data, and understood the concern, as well as being a FAIR expert themselves who could relay the concern accurately to the FAIR Expert team.

FAIR expert activities – building a transformation pipeline

"RDFizing" is the process of transforming a non-RDF data format into RDF. We tried two RDFizers that execute such transformations using RML as the mapping language – RMLMapper [21], and SDM-RDFizer [22]. RMLMapper has a rich set of features, including the ability to encode transformation rules that can trigger execution of algorithms over a CSV cell prior to the RDF transformation. SDM-RDFizer conversely, lacks these powerful extensions, but is significantly faster in our (informal) head-to-head tests. Since YARRRML currently cannot encode rules, we do not benefit from the additional power provided by RMLMapper, and thus selected SDM-RDFizer for this modelling initiative. Nevertheless, the choice of RDFizing technology can be revisited at a later date, without affecting any of our other decisions.

For storage of the resulting Linked Data, we have selected GraphDB [23], due to its ongoing support by the developers, the availability of a free (though not open source) version, and the availability of a fairly comprehensive API for mechanization of data loading, maintenance, and querying. GraphDB also supports access control methods which provide options for securing access to the FAIRified dataset. A "bootstrapping" Docker image for GraphDB was created to ensure that GraphDB is installed and configured correctly, thus eliminating the need for the registry host to have this expertise.

Deployment of the ETL pipeline is achieved via docker-compose, where every component has been "dockerized" and uses a Docker [24] network to facilitate communication between the components. This ensures that there are no unnecessary ports or APIs exposed on the registry server, helping maintain the security of their internal space. The three components mentioned above - RMLMapper, SDM-RDFizer, and GraphDB - are coordinated via a fourth Docker container, representing an orchestration tool. The orchestrator is triggered by a Web call to its interface. Once initiated, it automatically refreshes the current database of YARRRML templates from the CDE Project GitHub, and then examines the content of a folder shared with the host. This shared folder contains the host's CSV files that will be subject to RDF transformation. Using filename-matching, the system matches each CSV with an appropriate YARRRML template and executes the transformation. After all transformations have completed, a connection is opened to GraphDB, all previous data is deleted, and the refreshed data is uploaded.

The suite of docker images are referred-to as the "CDE-in-a-Box", and the instructions for running the bootstrapping process, as well as how to interact with CDE-in-a-Box, are available on a dedicated Git [25].

FAIR expert activities – testing

Speed tests were run by calling RMLMapper and SDM-RDFizer images via docker-compose on a Linux PC. A variety of exemplar 10.000 row CSV files and YARRRML templates were used for the measurement and execution process. The average speed of RDF triple generation was 12,500 triples per second. The tests were run on an AMD Ryzen 73800XT 3.9 GHz CPU workstation, with 32 Gb 3200Mhz RAM memory, RTX 2070 Super 8 Gb GPU and M.2. NVMe SSD memory. Quality-control tests will, largely, be registry-specific, though we are considering possible mechanisms for generalizing this problem through the use of Shape Expressions (ShEx) validation (described in "**Future Work**" in the Discussion section).

Results

The models

The models created to capture the 16 CDEs are described in Table 2, and are available in the CDE Project GitHub .

To help data custodians understand the models, they are generated and published in a variety of formats. Most importantly, an exemplar "runnable" CSV file, which is documented on a Web page – one page per CDE – containing a description of the CDE Model, its intended use, the CSV column headers, the constraints on the content of each column, and any usage notes that will assist the data custodians in their understanding of how to generate compatible CSV. A screenshot of the documentation is provided in Fig. 2.

To assist both data custodians and data consumers, a variety of other representations are also generated. When the exemplar CSV is run through the transformation pipeline, the resulting RDF file is then converted into a model image via a semi-automated mechanism [37]. A ShEx model is also created to allow data custodians (and users) to validate these transformations. The

Table 2 Models created to represent the CDEs. Models are created in YARRRML and made available on the CDE Project GitHub, accompanied by markdown documentation explaining the structure of an appropriate CSV file. Note that not all EU RD CDEs appear 1-to-1 with a CDE model. This is because, for example, the consent CDE can be reused for diverse types of consent (e.g., consent for contact, consent for data reuse), and the Pseudonym CDE is a part of every other model, and therefore has not been modelled as an independent element

CDE Model Name	Purpose
Disease Progression [26]	A "container" node to group together all other CDEs that refer to the same diagnosis. For example, the "age of diagnosis" CDE is related to a specific rare disease via traversal into the "disease progression" container, and then traversal into the "diagnosis" CDE that is also connected to "disease progression"
Care Pathway [27]	Captures the date of first contact with the specialist healthcare system; is connected to "disease progression"
Diagnosis [28]	Captures the final disease diagnosis using ORPHA codes; is connected to "disease progression"
Disease History [29]	Captures age at first symptoms and age at diagnosis; is connected to "disease progression"
Genetic Diagnosis [30]	Captures the sequence variant(s) found in this patient, using a variety of different coding systems; is connected to "disease progression"
Patient Consent [31]	Captures the consent of the patient over several axes (e.g., consent for contact, consent for data reuse). Provides a reference to the signed consent form, as well as an input reference to the (blank) consent template.
Patient Status [32]	Captures the current status of the patient, and their date of death if the patient is deceased
Personal Information [33]	Captures (superficial) personal information such as birth date and sex (there are ongoing debates in the EJP modelling group as to whether this should be converted to an age, or an age-range, for improved privacy)
Phenotyping [34]	Captures the phenotypes of the patient, using Human Phenotype Ontology terms
Disability [35]	Captures the score for a disability test. The specific test administered is indicated as one of the child nodes of obo: NCIT_C20993 (Clinical or Research Assessment Tool), and thus this CDE model is broadly useful for many disorders.
Undiagnosed [36]	Captures the case where a patient has phenotypic anomalies, and an identified sequence variant, but for some reason has not been definitively diagnosed.

ShEx models are manually created according to the Shape Expressions 2.1 Primer specification [38], and the resulting ShEx file is converted into an image via the RDFShape tool [39]. An exemplar RDF visualization for CDE #3 "Patient Status" is diagrammed in Fig. 3, and a diagram of the ShEx validator for that model is shown in Fig. 4.

Model filling - "CDE in a Box"

As described in the Methods section, the CDE-in-a-Box is deployed via docker-compose and is triggered by a simple Web call to a local address. Transformed data is automatically loaded into a CDE-specific data store on GraphDB, ensuring that the security constraints on this data can be managed independently of other datasets provided by the software.

The relationship of all of the components to one another, and the responsibilities of each participant, is diagrammed in Fig. 5.

Discussion

When undertaking any modelling activity, there is always the potential to "over-fit" the model. To this end, we have been attempting to apply the model to datatypes other than those covered in the CDE list. Specifically, we have looked at three very distinct datatypes: physical body measurements, laboratory tests, and Patient-Reported Outcome Measurements (PROMs), which are a questionnaire-style metric. In all cases, we were able to generate the Linked Data record with few or no changes to the core model. In particular, the Physical Body measurements required only an additional link to a measurement protocol; for PROMs we added an Input to the Process node representing the PROM question; and for Laboratory Tests we extended this further where an Input is included - constrained to being a body tissue - a "target" is included - constrained to being the compound being measured - and link is added to the measurement protocol document (see Fig. 6). Hence, we believe that this core model is capable of representing the majority of data entities we will encounter in the biomedical/clinical space with only minor modifications.

With respect to generalizability and scalability of this approach, a comprehensive survey of the European Reference Networks (ERNs) participating in EJP RD revealed 13 categories of data from 16 ERN data dictionaries; for example, "laboratory tests" and "personal information" are two such categories. Every category requires a YARRRML template to be constructed, following the core pattern but changing, for example, the default ontological types of each node, and the column header names. We have built code libraries that automatically generate these YARRRML templates via a simple API, and thus in practice, a new YARRRML model can be created in approximately one hour, now that the general pattern has been established.

Documentation of the model, and decisions about the constraints on the allowed content of each CSV column

Patient Status CDE

The CDE for patient status

CSV file

Example CSV file

Please find example CSV file here

Columns

pid, uniqid, date, status_uri, status_label, death_date

Notes:

- pid - patient unique identifier
- uniqid: some row unique identifier, over all sessions (a millisecond timestamp, numerical only, is a good idea)
- date: ISO 8601 formatted date representing the date of the current observation
- status_uri: one of
  - http://semanticscience.org/resource/SIO_010059 (dead)
  - http://semanticscience.org/resource/SIO_010058 (alive)
  - http://purl.obolibrary.org/obo/NCIT_C70740 (lost to follow-up)
  - http://purl.obolibrary.org/obo/NCIT_C124784 (refused to participate)
- status label: a human readable label to match the value of the status URI for that row
- death_date: if the patient is dead, the recorded date of death (may be different from the 'date' column of this record). If patient is not dead, leave this field as empty

YARRRML

Please find the YARRRML template for this module here

Fig. 2 The Markdown documentation explaining how to prepare a CSV file for the "Patient Status" CDE. Documentation includes, where appropriate, the restrictions on the possible values in a given column, such as 'status uri' in this example

takes more consideration and time, though all elements of a well-documented model can be easily created in less than a day. This, however, leads to a problem for which the correct solution is, as yet, not known. Because the models themselves are generalized, the problem of selecting the correct specific value for a given column becomes a task for the data provider. For example, in the Body Measurement model, we document that the "attribute being measured" column should contain an ontology URI that is a child of obo: NCIT_C19332 (personal attribute). While many of the participating hosts are familiar with ontologies, "coding" (the act of assigning a controlled vocabulary term to a concept, observation, or phenomenon) is an activity primarily undertaken by insurance and governmental organizations, and by trained disease classifiers, and as such many other participants will not have this experience. Thus, we suspect that this task may be difficult for a subset of our registry participants. One alternative is that the FAIR Experts create a specific model for every case (every attribute, every lab measurement, etc.). This, however, would result in many highly specific models, and would in turn, require the data host to generate separate CSV files for each model. The alternative is to keep the models generic and find another way to provide advice or support to the data hosts as they generate the CSV. We are exploring both solutions to gain a better understanding of how to address this problem in the future.

The transformation step itself – generating ~12,500 RDF statements per second – would appear to be sufficiently fast that it would be possible to generate a new snapshot of a registry on a nightly basis.

Finally, the models are intended to be reusable, and the onus of creating a matching CSV is put on the data custodians/experts. Another approach would have been to create a comprehensive transformation map of the

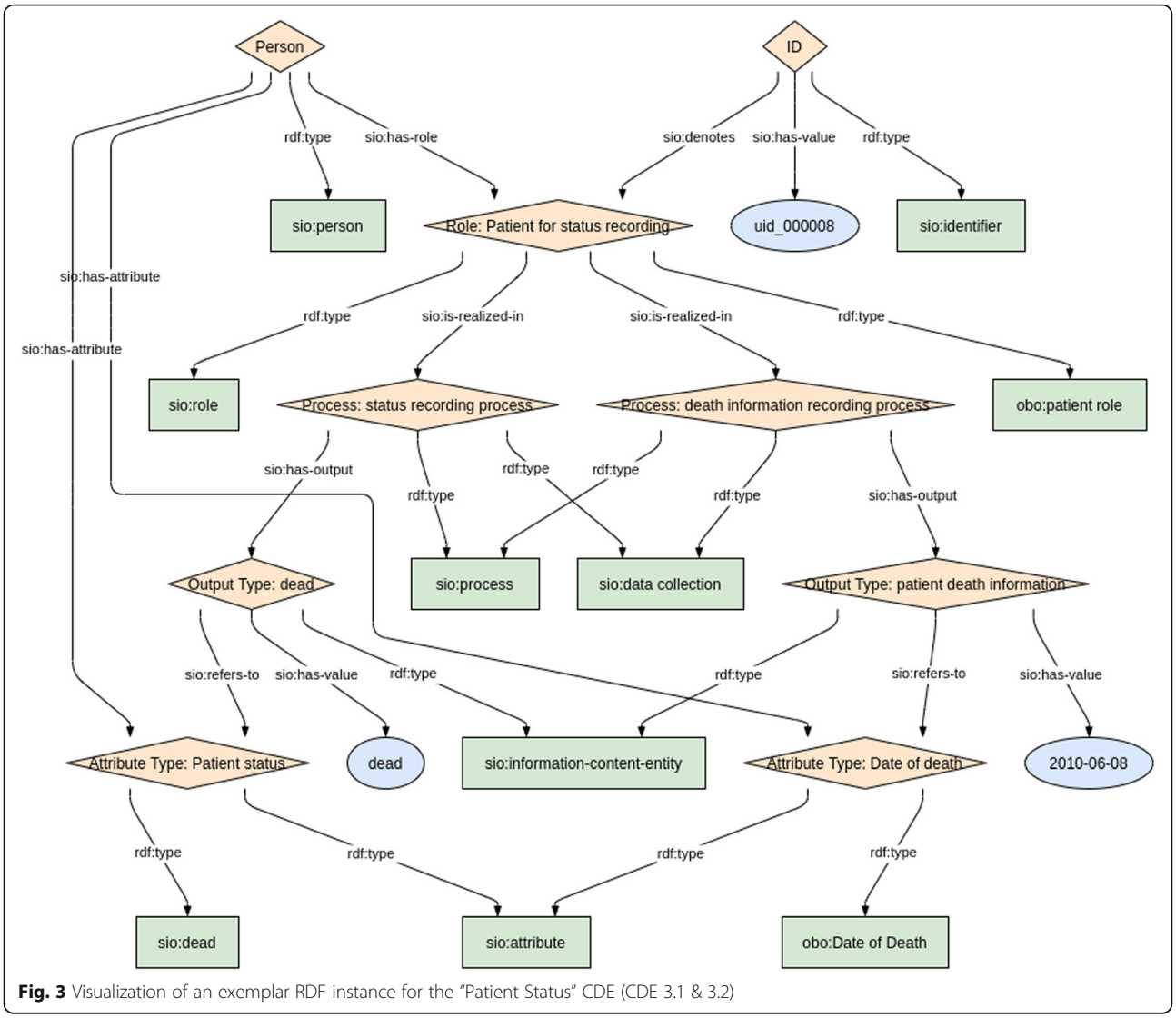


Fig. 3 Visualization of an exemplar RDF instance for the "Patient Status" CDE (CDE 3.1 & 3.2)

entire dataset, for every participating registry using, for example, R2RML [40]. Since the data custodians are not anticipated to be FAIR experts, this approach would have centralized the problem of mapping into the hands of the few core technologists in the EJP RD, which we feel is a much less practical solution, and less scalable. The solution proposed here distributes the effort over many more participants, and the sharing of a set of core models ensures that, despite being non-coordinating, the participating registries will nevertheless generate interoperable outputs.

Exemplar use-case

Two registries about vascular rare diseases have entered into a data sharing agreement for a study on the relationship between identical mutations and phenotype/disability over many individuals. They select the Genetic Diagnosis, Phenotyping and Disability CDE as those that will contain the most relevant data. One registry executes three SQL queries on their Oracle database, which generate three CSV files following the CDE Model Templates. They activate CDE-in-a-Box, which converts those data into FAIR Data loads it into a database within their own secure space. The partner has their data in the form of a series of MS Excel spreadsheets. They export from spreadsheet into CSV, and similarly activate CDE-in-a-box to generate FAIR Data. The investigative query is shared by both sites, and executed within their secure spaces, where they exchange only the query results. They are confident that they are extracting and integrating the full gamut of information from both sites because of the harmonization of structural and ontological choices that are enforced by the CDE models, yet their individual tasks only involved generating the CSV and executing the query. Thus, the process of

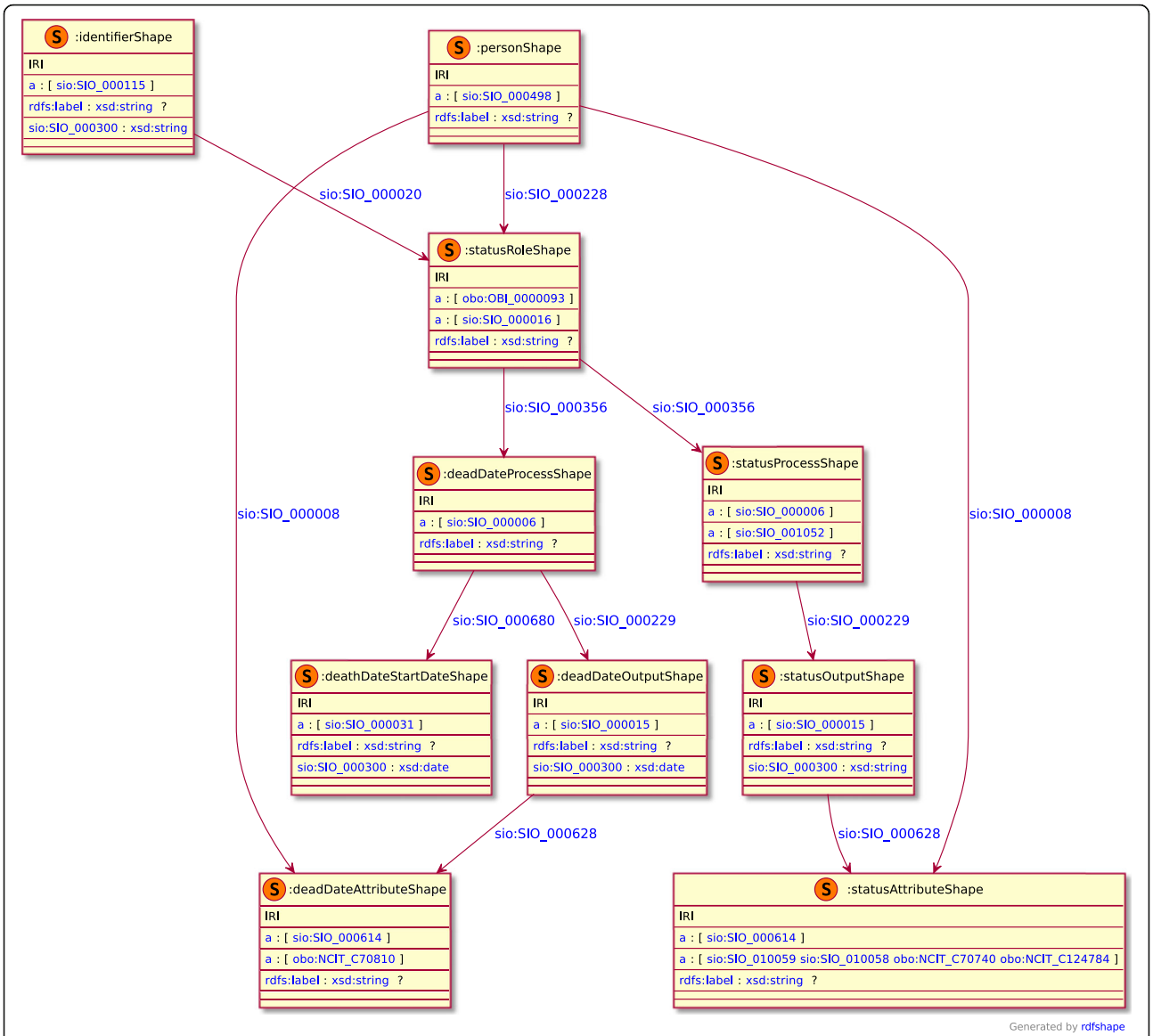


Fig. 4 Visualization of the ShEx validation shape for the Patient Status CDE data

querying over both datasets was quite straightforward for both participants, despite having notably different underlying data structures.

Peer initiatives

Beyond CDEs, the most widely used health care data exchange formats are all exploring FAIR-oriented mappings. OMOP CDM [41] and ContSys [42] were recently compared [43] for their ability to be transformed to one another, and to have their facet values captured in (largely) SNOMED CT to enhance their FAIRness. The HL7's FHIR4FAIR project began its public facing activities at a "connectathon" event in early 2021 [44] and expects to have an early and final normative document in late 2023 and early 2024, respectively describing (among other things) how to apply the RDA FAIR Maturity Model Working Group Maturity Indicators [45] to FHIR data structures (both manual and automated) and define a minimal metadata set for health recordsets. The Critical Path Institute (C-PATH) [46] is working in parallel with the FAIR data transformation subgroup within the EJP RD to attempt to achieve a mapping between the Clinical Data Interchange Standards Consortium (CDISC) [47] standards and the EJP RD semantic CDEs, with their first attempt being to use LinkML/BioLink [48] as a domain-neutral abstract representation of the clinical data that might act as a "Rosetta stone". Finally, the OpenEHR initiative has adopted design principles [49] that enables computable semantics in their data models [50]. Thus, the latter of these peer initiatives resembles

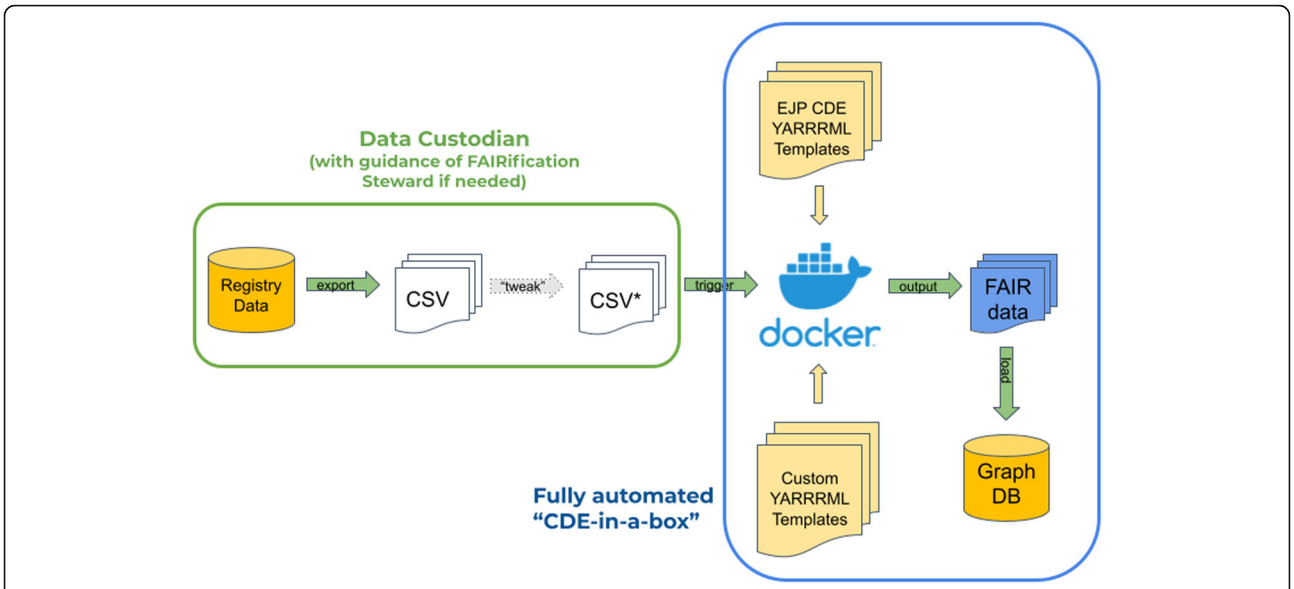


Fig. 5 The components of the workflow annotated with the responsibilities of the parties. The left side of the diagram, outlined in green, are the responsibilities of the data custodian in collaboration with the Data Steward. This includes export of their registry data into CSV format, and possibly some additional modification of that exported data to conform to the template. On the right is the fully automated CDE-in-a-Box, which is constructed by the FAIR Expert team and provided as a docker-compose installation. The arrow labelled "trigger" is the Web page call that the data custodian makes when they are ready to execute their transformation

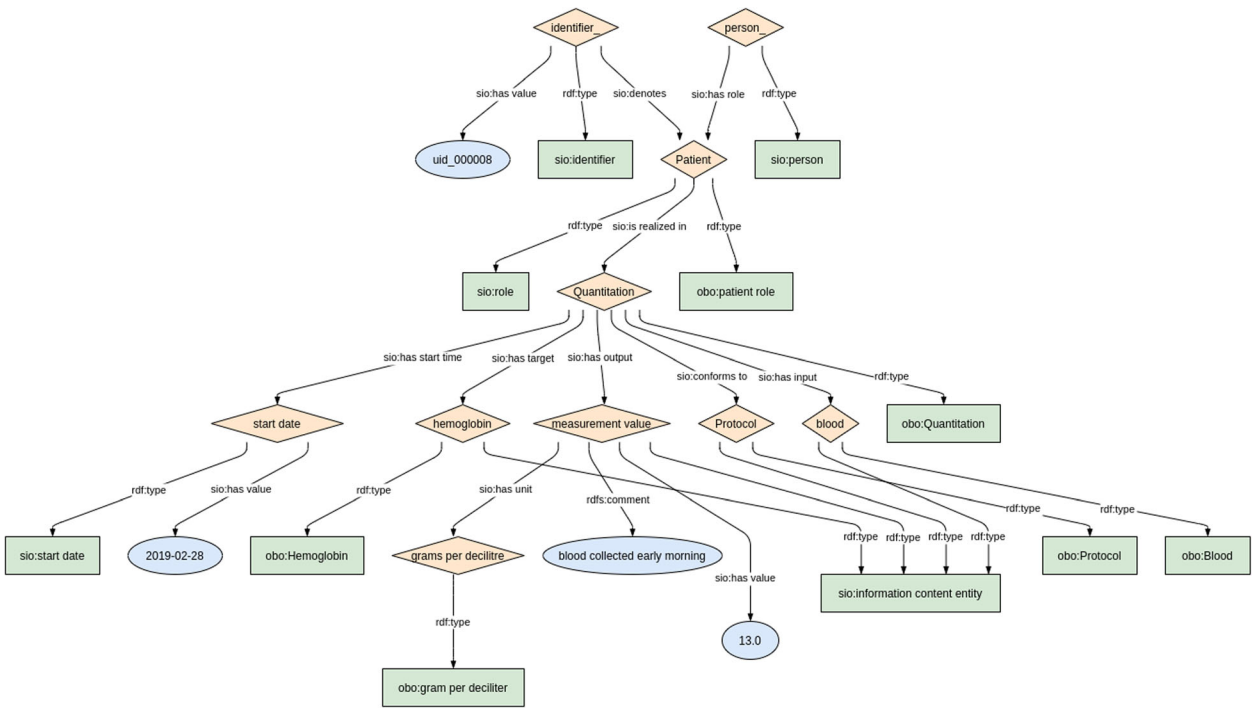


Fig. 6 The model for Laboratory Measurements. Of note are the three new connections on the "Quantitation" (Process) node – one representing the input (blood), one representing the target molecule (haemoglobin), and the third representing the link to the protocol. The remainder of the model is (structurally) identical to the core model shown in Fig. 1

our own attempts to build an overall model based in rich semantics; however, we differ from most of these peer initiatives in that we are attempting to map arbitrary existing formats from non-coordinating registries into a unified, semantically-grounded (in SIO) FAIR model, versus taking an existing standard model and attempting to make it more FAIR.

A notable peer initiative, with highly similar goals to our own, is found in PennTURBO [51], which also pursues a multi-step transformation process to achieve a final rich semantic model. PennTURBO's workflow mirrors ours in several ways: first, there is a transformation from native format into a semantically-impoverished intermediate representation (in our case, CSV, and in PennTURBO, a graph). They then similarly use a domain specific language to transform the intermediate representation into the final semantically rich model (in our case, RML, and in PennTURBO, SPARQL). A key distinction between the two projects is the point of responsibility for the generation of the intermediate format. PennTURBO includes a set of data extraction/transformation instructions for each data source, thus the responsibility for the correct interpretation of the data source, and its export, is held by the PennTURBO development team. In contrast, the transformation pipeline presented here creates only a richly documented intermediate template, which must then be filled by any data source that wishes to participate. Thus, the responsibility for correctly filling that template is pushed to the participant themselves, as is the correct interpretation of the data holdings. It is an open question whether the decentralized approach we have taken will be more scalable than PennTURBO, or if the same decentralization will lead to more erroneous template filling, as the responsibility for accurate data export moves further away from the FAIR experts. Nevertheless, like PennTURBO, our mapping extends beyond identifying appropriate ontology terms for each data facet, and both projects share a goal of attempting to better model the activities around the creation of data – creating a "digital twin" for the data, which as a beneficial consequence, provides model positions for metadata about every element, including the participants, the relationships between them, and the process' protocol and other annotations.

Future work

Work is underway to automate the creation of ShEx models for all CDEs, and use them to add a quality-checking layer into the transformation pipeline. Moreover, we additionally plan to use ShEx to publish a public model of the entire contained dataset, which we believe can be used both to aid discovery, but more importantly, to facilitate future efforts around federated queries.

With the goal of allowing future extension of these models – for example, by expanding the ontological concepts allowed as possible values, or adding new or repository-specific metadata we will soon begin to provide training to those who wish to learn how to build or edit the YARRRML templates themselves. We are improving the tooling that facilitates construction of these templates to better enable registry custodians to expand or diversify the templates without necessarily requesting help from the EJP RD modelling team. In this way, we hope that the core data will be interoperable, even if individual sites add enhanced metadata that is not in-common with other registries. Moreover, dissemination of the expertise around template-building provides a path to self-sustainability of this initiative, beyond the end of the EJP-RD project.

Extension and revision, however, must be done with a recognition of why this centralized modelling initiative was deemed necessary. Interoperability is difficult to achieve, particularly without agreement on the concepts being modelled. Moreover, the decision to use a model backbone with very strict semantics (SIO) makes it necessary to be extremely careful in the selection of ontology terms – ensuring, for example, that there is a distinction between the concepts of "blood pressure" as a quality/attribute of a patient, versus "blood pressure" referring to the output of a measuring process. These kinds of decisions require expertise and experience in ontology construction and use. As such, extension of the models in a distributed manner by end-users introduces several risks regarding interoperability, including lacking mappings between ontologies, reduced shared semantics, and restricted use of ontologies, for example due to licensing restrictions. The same issue has been highlighted in other CDE initiatives, and was addressed in a recent overview of the problems related to CDE mapping [52] (using the term in its most general sense, not specifically the RD CDEs that are discussed in this manuscript). They noted that the objective of CDEs – to assist in the harmonization of data between independent studies – was being thwarted by imprecise definitions of those CDEs (a problem shared with the CDEs upon which this study is based). They further noted that "*CDEs can deliver more value when they conform to accepted data standards, are bound to terminologies and are used consistently across studies*", and that, for this reason, CDE-focused initiatives are falling far short of the objectives of FAIRness.

Conclusions

We undertook a process of constructing a reusable, generic data model, based on the design principles of the Semantic Science Integrated Ontology, to represent all the EU Rare Disease Platform Common Data Elements.

Emergent mapping technologies such as YARRRML, RML, and "RDFizing" tools allowed us to create an automated pipeline for filling these data models starting from a well-documented CSV template – a format accessible to all our target end-users. We demonstrated the generic nature of the model by successfully extending it - while remaining within the overall architecture of SIO - to widely disparate non-CDE clinical data within the Rare Disease space. Feedback from end-users indicates that they found this solution helpful, and easy to apply. As FAIR data publishing becomes increasingly an expectation – even a requirement – of funding agencies and publishers, there is an urgent need for straightforward tooling to assist data providers to comply with these expectations. In many cases, those who generate data will not have expertise in data modelling, and particularly not in semantically grounded data modelling, as is a requirement of FAIR. The activities and workflows described here indicate that the approach of building a generic, reusable, models, and an automated pipeline to fill them, will be widely applicable in biomedicine and beyond.

Abbreviations

API: Application Programming Interface; BYODs: Bring Your Own Data workshops; CDE: Common Data Element; CSV: Comma Separated Values; EJP RD: European Joint Programme on Rare Disease; ERN: European Reference Network; ETL: Extract, Transform, Load; FAIR: Findable, Accessible, Interoperable, Reusable; GUID: Globally Unique Identifier; OMIM: Online Mendelian Inheritance in Man; OWL: Web Ontology Language; PROM: Patient Reported Outcome Measures; PROV-O: Provenance Ontology; RD: Rare disease; RDF: Resource Description Framework; RML: RDF Mapping Language; ShEx: Shape Expression; SIO: SemanticScience Integrated Ontology; URI: Uniform Resource Identifier; YAML: YAML Ain't Markup Language

Acknowledgements

We would like to express our gratitude to the Duchenne Parent Project in the Netherlands for allowing the code for portions of their FAIR transformation solution to be open source and included in the CDE in a Box. This solution was specifically commissioned and developed for their patient-led registry - The Duchenne Data Platform - which had undergone a FAIRification process in 2021. Their generosity stems from their continuous belief in FAIR as a new paradigm for optimising data visiting and analysis and as a result, pledged to support others in their own FAIR data endeavours. We thank Foundation 29, the software developers of the Duchenne Data Platform, for their technical expertise and positive collaboration during the development of the Duchenne FAIR project. We furthermore wish to thank Leo Schultze Kool for helping us start the CDE modelling process for the FAIRification of the VASCA registry, his support for FAIR implementation of registries, and valuable feedback on the interpretation of CDEs from a clinical perspective, and Peter-Bram 't Hoen for his continuous active support of our efforts. Finally, the authors acknowledge the support of Ana Rath and Franz Schaefer in the development of a conceptual framework for the EJP RD Virtual Platform.

Authors' contributions

All authors contributed to this work through participation in model design and design workshops, weekly discussions, and writing and revising this manuscript. Specific additional contributions are as follows: MDW & RK created CDE-in-a-Box code and docker images; RK created workflow and code to generate RDF model images; MDW created YARRRML template files, sample CSV files, and markdown documentation; MD provided guidance on SIO modelling, expanded SIO design patterns and edited the SIO ontology; RC & MR cross-referenced CDE models with other clinical data models and checked all semantics for the modelling group; PA executed quality control of workflow outputs, creation and review of images for RDF and ShEx models; BdSV reviewed user instructions, updated documentation, tested CDE in a box Platform; NB planned and hosted "designathons" and other workshops, and gathered feedback; LJSK co-authored the first draft of the CDE model; PvD reviewed parts of the model for ontological consistency; NvL and the Duchenne Parent Project initiated the code development project that led to the CDE in a Box transformation pipeline, and provided end-user review and feedback; SZ reviewed RDF and ShEx files; BdSV, SZ, CHB, JKvdV, CLC executed surveys of registries and collected feedback from end-users. The author(s) read and approved the final manuscript.

Funding

All authors with the exception of MD are supported by the funding from the European Union's Horizon 2020 research and innovation programme under the EJP RD COFUND-EJP N° 825575. Funding for MD is provided by the European Union's Horizon 2020 research and innovation programme under grant agreement N° 824087.

Availability of data and materials

Project name: CDE Semantic Model Implementations.
Description: Parent GitHub repository for individual sub-components that generate the YARRRML templates; The YARRRML templates themselves, plus sample CSV data and CSV documentation; and the docker-compose repository used by "CDE in a Box".
Project home page: https://github.com/ejp-rd-vp/CDE-semantic-model-implementations
Operating system(s): Platform independent.
License: MIT (all subcomponents).
Project name: CDE Semantic Model.
Description: Diagrams, ShEx, and sample RDF for each of the Semantic Models.
Project home page: https://github.com/ejp-rd-vp/CDE-semantic-model
Operating system(s): Platform independent.
License: MIT.
Project name: CDE in a Box.
Description: Bootstrap and Production docker-compose files to run CDE in a Box.
Project home page: https://github.com/ejp-rd-vp/cde-in-box
Operating system(s): Platform independent.
License: Apache 2.0.
Project name: FAIR in a Box.
Description: A fork from CDE in a Box (above) that takes a different approach to installing and orchestrating the docker images.
Project home page: https://github.com/markwilkinson/FAIR-in-a-box
Operating system(s): Platform independent.
License: Apache 2.0.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Leiden University Medical Center, Leiden, The Netherlands. [2]Departamento de Biotecnología-Biología Vegetal, Escuela Técnica Superior de Ingeniería Agronómica, Alimentaria y de Biosistemas, Centro de Biotecnología y Genómica de Plantas (CBGP), Universidad Politécnica de Madrid (UPM), Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA), Pozuelo de Alarcón, Madrid, ES, Spain. [3]Department of Medical Informatics, Amsterdam Public Health Research Institute, Amsterdam UMC, University of Amsterdam, Meibergdreef 9, Amsterdam, The Netherlands. [4]Department of Medical Imaging, Radboud University Medical Center, Nijmegen, The Netherlands. [5]Centre for Molecular and Biomolecular Informatics, Radboud

University Medical Center, Nijmegen, The Netherlands. [6]Institute of Data Science, Paul-Henri Spaaklaan 1, Maastricht University, 6229EN Maastricht, The Netherlands. [7]Division of Paediatric Nephrology, Centre for Paediatrics and Adolescent Medicine, University of Heidelberg, Heidelberg, Germany. [8]University of Groningen and University Medical Center Groningen, Genomics Coordination Center and Department of Genetics, Antonius Deusinglaan 1, 9713, AV, Groningen, The Netherlands. [9]Duchenne Parent Project, Veenendaal, The Netherlands.

Received: 26 July 2021 Accepted: 23 February 2022
Published online: 15 March 2022

References

1. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. Comment: The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. England. 2016;3:160018.
2. Set of Common Data Elements | EU RD Platform [Internet]. [cited 2021 Jul 8]. Available from: https://eu-rd-platform.jrc.ec.europa.eu/set-of-common-data-elements_en
3. Lopes P, Roos M. Bring your own data parties and beyond: make your data linkable to speed up rare disease research. Rare Dis Orphan Drugs [Internet]. 2014;1:21–4 Available from: http://rarejournal.org/index.php/rarejournal/article/download/69/93.
4. Roos M, Gray A, Waagmeester A, Thompson M, Kaliyaperumal R, Horst EVD, et al. Bring Your Own Data Workshops: A Mechanism to Aid Data Owners to Comply with Linked Data Best Practices. Proc 7th Int Work Semant Web Appl Tools Life Sci (SWAT4LS 2014). 2014;
5. DCMI: DCMI Metadata Terms [Internet]. [cited 2021 Dec 14]. Available from: https://www.dublincore.org/specifications/dublin-core/dcmi-terms/
6. Data Catalog Vocabulary (DCAT) - Version 2 [Internet]. [cited 2021 Dec 14]. Available from: https://www.w3.org/TR/vocab-dcat-2/
7. Hooft R, Goble C, Evelo C, Roos M, Sansone S, Ehrhart F, et al. ELIXIR-EXCELERATE D5.3: Bring Your Own Data (BYOD). 2019 [cited 2021 Jul 8]; Available from: https://zenodo.org/record/3207809
8. Jacobsen A, Waagmeester A, Kaliyaperumal R, Stupp GS, M. Schriml L, Thompson M, et al. Wikidata as an intuitive resource towards semantic data modeling in data FAIRification. Semantic Web Applications and Tools for Healthcare and Life Sciences; 2018 [cited 2021 Dec 16]; Available from: /articles/journal_contribution/Wikidata_as_an_intuitive_resource_towards_semantic_data_modeling_in_data_FAIRification/7415282/2.
9. LUMC-BioSemantics/ERN-common-data-elements [Internet]. [cited 2021 Jul 9]. Available from: https://github.com/LUMC-BioSemantics/ERN-common-data-elements
10. Kersloot MG, Jacobsen A, Groenen KHJ, dos Santos VB, Kaliyaperumal R, Abu-Hanna A, et al. De-novo FAIRification via an Electronic Data Capture system by automated transformation of filled electronic Case Report Forms into machine-readable data. J Biomed Inform. Academic Press. 2021;122:103897.
11. Groenen KHJ, Jacobsen A, Kersloot MG, dos Santos Vieira B, van Enckevort E, Kaliyaperumal R, et al. The de novo FAIRification process of a registry for vascular anomalies. Orphanet J Rare Dis [Internet]. BioMed Central Ltd; 2021 [cited 2021 Dec 16];16:1–10. Available from: https://doi.org/10.1186/s13023-021-02004-y
12. McKusick-Nathans Institute of Genetic Medicine. OMIM - Online Mendelian Inheritance in Man [Internet]. Johns Hopkins Univ. Baltimore. [cited 2021 Jul 6]. Available from: https://www.omim.org/
13. Köhler S, Gargano M, Matentzoglu N, Carmody LC, Lewis-Smith D, Vasilevsky NA, et al. The human phenotype ontology in 2021. Nucleic Acids Res [Internet]. Oxford University Press; 2021 [cited 2021 Jul 6];49:D1207–D1217. Available from: https://pubmed.ncbi.nlm.nih.gov/33264411/
14. Lebo, T, Sahoo, S, McGuinness, D, Belhajjame, K, Cheney, J, Corsar, D, Garijo, D, Soiland-Reyes, S, Zednik, S & Zhao J. PROV-O: The PROV Ontology [Internet]. W3C Recomm. World Wide Web Consort. 2013 [cited 2021 Jul 6]. Available from: https://www.w3.org/TR/prov-o/
15. Dumontier M, Baker CJO, Baran J, Callahan A, Chepelev L, Cruz-Toledo J, et al. The semanticscience integrated ontology (SIO) for biomedical research and knowledge discovery. J Biomed Semantics [Internet]. BioMed Central Ltd.; 2014 [cited 2021 Jul 6];5:1–11. Available from: http://sio.semanticscience.org.
16. Design Patterns · MaastrichtU-IDS/semanticscience Wiki [Internet]. [cited 2021 Jul 6]. Available from: https://github.com/MaastrichtU-IDS/semanticscience/wiki/Design-Patterns
17. Golbeck J, Fragoso G, Hartel F, Hendler J, Oberthaler J, Parsia B. The National Cancer Institute's Thesaurus and Ontology. SSRN Electron J [Internet]. Elsevier BV; 2003 [cited 2021 Dec 14]; Available from: https://papers.ssrn.com/abstract=3199007
18. Castor EDC [Internet]. Castor Electron. Data Capture. [cited 2021 Jul 6]. Available from: https://www.castoredc.com/
19. Heyvaert P, De Meester B, Dimou A, Verborgh R. Declarative rules for linked data generation at your fingertips! Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) [Internet]. Springer Verlag; 2018 [cited 2021 Jul 6]. p. 213–7. Available from: https://doi.org/10.1007/978-3-319-98192-5_40
20. CDE-semantic-model-implementations: This repository contains implementation artifacts related to CDE-semantic-model [Internet]. [cited 2022 Feb 8]. Available from: https://github.com/ejp-rd-vp/CDE-semantic-model-implementations
21. RMLio/RML-Mapper: Generate High Quality Linked Data from multiple originally (semi-)structured data (legacy) [Internet]. [cited 2021 Jul 6]. Available from: https://github.com/RMLio/RML-Mapper
22. Iglesias E, Jozashoori S, Chaves-Fraga D, Collarana D, Vidal ME. SDM-RDFizer: An RML Interpreter for the Efficient Creation of RDF Knowledge Graphs. Int Conf Inf Knowl Manag Proc. Association for Computing Machinery. 2020: 3039–46.
23. GraphDB Downloads and Resources [Internet]. [cited 2021 Jul 6]. Available from: https://graphdb.ontotext.com/
24. Empowering App Development for Developers | Docker [Internet]. [cited 2021 Jul 6]. Available from: https://www.docker.com/
25. CDE-in-box: This repository contains software to create and deploy CDEs [Internet]. [cited 2021 Jul 6]. Available from: https://github.com/ejp-rd-vp/cde-in-box
26. CDE-semantic-model-implementations/disease_progression_csv_template.md [Internet]. [cited 2021 Jul 6]. Available from: https://github.com/ejp-rd-vp/CDE-semantic-model-implementations/blob/master/YARRRML_Transform_Templates/docs/disease_progression_csv_template.md
27. CDE-semantic-model-implementations/care_pathway_csv_template.md [Internet]. [cited 2021 Jul 6]. Available from: https://github.com/ejp-rd-vp/CDE-semantic-model-implementations/blob/master/YARRRML_Transform_Templates/docs/care_pathway_csv_template.md
28. CDE-semantic-model-implementations/diagnosis_csv_template.md [Internet]. [cited 2021 Jul 6]. Available from: https://github.com/ejp-rd-vp/CDE-semantic-model-implementations/blob/master/YARRRML_Transform_Templates/docs/diagnosis_csv_template.md
29. CDE-semantic-model-implementations/disease_history_csv_template.md [Internet]. [cited 2021 Jul 6]. Available from: https://github.com/ejp-rd-vp/CDE-semantic-model-implementations/blob/master/YARRRML_Transform_Templates/docs/disease_history_csv_template.md
30. CDE-semantic-model-implementations/genetic_diagnosis_csv_template.md [Internet]. [cited 2021 Jul 6]. Available from: https://github.com/ejp-rd-vp/CDE-semantic-model-implementations/blob/master/YARRRML_Transform_Templates/docs/genetic_diagnosis_csv_template.md
31. CDE-semantic-model-implementations/patient_consent_csv_template.md [Internet]. [cited 2021 Jul 6]. Available from: https://github.com/ejp-rd-vp/CDE-semantic-model-implementations/blob/master/YARRRML_Transform_Templates/docs/patient_consent_csv_template.md
32. CDE-semantic-model-implementations/patient_status_csv_template.md [Internet]. [cited 2021 Jul 6]. Available from: https://github.com/ejp-rd-vp/CDE-semantic-model-implementations/blob/master/YARRRML_Transform_Templates/docs/patient_status_csv_template.md
33. CDE-semantic-model-implementations/personal_csv_template.md [Internet]. [cited 2021 Jul 6]. Available from: https://github.com/ejp-rd-vp/CDE-semantic-model-implementations/blob/master/YARRRML_Transform_Templates/docs/personal_csv_template.md
34. CDE-semantic-model-implementations/phenotyping_csv_template.md [Internet]. [cited 2021 Jul 6]. Available from: https://github.com/ejp-rd-vp/CDE-semantic-model-implementations/blob/master/YARRRML_Transform_Templates/docs/phenotyping_csv_template.md
35. CDE-semantic-model-implementations/disability_csv_template.md [Internet]. [cited 2021 Jul 6]. Available from: https://github.com/ejp-rd-vp/CDE-semantic-model-implementations/blob/master/YARRRML_Transform_Templates/docs/disability_csv_template.md

36. CDE-semantic-model-implementations/undiagnosed_csv_template.md [Internet]. [cited 2021 Jul 6]. Available from: https://github.com/ejp-rd-vp/CDE-semantic-model-implementations/blob/master/YARRRML_Transform_Templates/docs/undiagnosed_csv_template.md
37. Draw.io RDF drawing utils api alpha version [Internet]. [cited 2021 Jul 6]. Available from: http://ejprd.fair-dtls.surf-hosted.nl:5000/
38. Shape Expressions (ShEx) 2.1 Primer [Internet]. [cited 2021 Jul 6]. Available from: http://shex.io/shex-primer/index.html
39. RDFShape — Playground for RDF, ShEx, SHACL and more [Internet]. [cited 2021 Jul 6]. Available from: https://rdfshape.weso.es/
40. R2RML: RDB to RDF Mapping Language [Internet]. [cited 2021 Jul 6]. Available from: https://www.w3.org/TR/r2rml/
41. OMOP Common Data Model – OHDSI [Internet]. [cited 2021 Dec 14]. Available from: https://www.ohdsi.org/data-standardization/the-common-data-model/
42. A system of concepts for the continuity of care [Internet]. [cited 2021 Dec 14]. Available from: https://contsys.org/page/default.
43. de Groot R, Cornet R, de Keizer N, Benis N, Raiez F. OMOP CDM compared to ContSys (ISO13940) to make data FAIR [Internet]. [cited 2022 Mar 8]. Available from: https://www.ohdsi.org/2020-global-symposium-showcase-52/.
44. HL7 FHIR FHIR4FAIR IG PSS - Services Oriented Architecture - Confluence [Internet]. [cited 2021 Dec 14]. Available from: https://confluence.hl7.org/display/SOA/HL7+FHIR+FHIR4FAIR+IG+PSS
45. FAIR Data Maturity Model Working Group. FAIR Data Maturity Model. Specification and Guidelines. 2020 [cited 2021 Dec 14]; Available from: https://zenodo.org/record/3909563
46. Critical Path Institute [Internet]. [cited 2021 Dec 15]. Available from: https://c-path.org/
47. CDISC | Clear Data. Clear Impact. [Internet]. [cited 2021 Dec 14]. Available from: https://www.cdisc.org/
48. LinkML specification — linkml documentation [Internet]. [cited 2021 Dec 14]. Available from: https://linkml.io/linkml/specifications/linkml-spec.html
49. Bönisch C, Sargeant A, Wulff A, Parciak M, Bauer CR, Sax U. FAIRness of openEHR Archetypes and Templates. SWAT4HCLS [Internet]. 2019 [cited 2021 Dec 14]. p. 102–11. Available from: https://www.openehr.org/ckm/
50. Frexia F, Mascia C, Lianas L, Delussu G, Sulis A, Meloni V, et al. openEHR Is FAIR-Enabling by Design. Public Heal Informatics Proc MIE 2021 [Internet]. IOS Press; 2021 [cited 2021 Dec 14];113–7. Available from: https://doi.org/10.3233/SHTI210131
51. Freedman HG, Williams H, Miller MA, Birtwell D, Mowery DL, Stoeckert CJ. A novel tool for standardizing clinical data in a semantically rich model. J Biomed Inform. Academic Press. 2020;112:100086.
52. Kush RD, Warzel D, Kush MA, Sherman A, Navarro EA, Fitzmartin R, et al. FAIR data sharing: The roles of common data elements and harmonization. J Biomed Inform. Academic Press. 2020;107:103421.

Publisher's Note

Journal of the Society for Clinical Data Management

Alarcón-Moreno P, et al. Leveraging Biolink as a FAIR "Rosetta Stone" Between Clinical Semantic Models Provides Emergent Interoperability. *Journal of the Society for Clinical Data Management*. 2022; 2(3): 2, pp. 1–8. DOI: https://doi.org/10.47912/jscdm.130

ORIGINAL RESEARCH

Leveraging Biolink as a FAIR "Rosetta Stone" Between Clinical Semantic Models Provides Emergent Interoperability

Pablo Alarcón-Moreno*, Ian Braun†, Emily Hartley†, Daniel Olson†, Nirupama Benis‡, Ronald Cornet‡, Mark D. Wilkinson* and Ramona L. Walls†

Interoperability between clinical datasets is challenging due to, in part, the number of data models and vocabularies in use and the variety of implementations. Here we describe the first steps in an ongoing effort to achieve interoperability between two clinical datasets currently being constructed within independent international projects. Both are utilizing the FAIR Principles but have constructed their data models independently and have selected different ontologies. In this initial exploratory experiment, we examined the degree to which a mapping of both models into an independent schema, Biolink, can increase interoperability. Mapping was achieved by categorizing the key nodes in both data models as "types" of concepts in the Biolink schema. We found that with this very thin mapping in place, and without changing either model, queries could be constructed that extracted data from both datasets, demonstrating that at least some degree of interoperability had been achieved. Our results support the use of FAIR-compliant data representations, which are, by nature, more interoperable than legacy clinical data representations, even when the models have not been coordinated upfront.

Keywords: Common Data Element (C19984); Interoperability (C142381); Data Integration; FAIR; SDTM

Introduction

The achievement of personalized and precision medicine demands not only massive multi-modal analysis of medical records, but also detailed international healthcare data to build accurate predictive models properly stratified to different sub-populations. Data integration is particularly crucial for rare diseases, where data are scarce, extremely heterogeneous, and disseminated in many repositories around the globe. In this study, we tested how well the use of a high-level semantic model, Biolink,[1] was able to support integration of data from two independent sources: Critical Path Institute's (C-Path) Rare Disease Cures Accelerator Data and Analytics Platform (RDCA-DAP) and European Joint Programme on Rare Diseases (EJP-RD)—when both sources used Findable, Accessible, Interoperable, and Reusable (FAIR) Data Principles to represent their data, but each used an independently developed data model.[2]

Background

Healthcare data, like most scientific data, is spread over many formats and repositories. This alone makes them challenging to integrate, but the data also have features, such as being highly privacy-sensitive, that dramatically inhibit integration. Within life and health sciences, interoperability projects that focused on machine-actionability began in the late 1990s, and some came to fruition in the early 2000s. Some approaches pushed the interoperability problem on to the data owner, trying to enforce harmonization at-source, such as caBIO and TAPIR.[3–5] caBIO created an interface standard for the cancer genetics/genomics community that all participating organizations should implement; whereas TAPIR, for the biodiversity community, pursued a query language that all sites should respond to. With respect to more general-purpose approaches, myGrid and BioMoby were both Web-services-based interoperability projects that used semantic annotations of Web service interfaces,

* Departamento de Biotecnología-Biología Vegetal, Escuela Técnica Superior de Ingeniería Agronómica, Alimentaria y de Biosistemas and Centro de Biotecnología y Genómica de Plantas (CBGP, UPM-INIA), Universidad Politécnica de Madrid (UPM) – Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA), ES

† Data Collaboration Center, Critical Path Institute, 1730 E. River Rd., Tucson, AZ 85718, US

‡ Department of Medical Informatics, Amsterdam Public Health Research Institute, Amsterdam University Medical Centers, University of Amsterdam, Meibergdreef 9, Amsterdam, NL

Corresponding authors: Pablo Alarcón-Moreno (pabloalarconmoreno@gmail.com); Ian Braun (ibraun@c-path.org)

and in the case of BioMoby, an ontology-driven eXtensible Markup Language (XML) Schema to harmonize data structures.[6–8] SADI and SSWAP used the Semantic Web technologies of Resource Description Framework (RDF) and Web Ontology Language (OWL) to annotate Web Service interface definitions and required that all data passed between their participating services be represented as ontologically grounded RDF.[9–10] Notably, none of the approaches mentioned thus far have attempted to achieve integration purely at the level of the data itself. This is at least in part due to legacy data formats being either unstructured or structured (for the purpose of sharing) in the form of XML documents, XML being described as "far and away the most complex data model ever proposed" and "seriously flawed".[11]

The now more widespread use of Semantic Web technologies should, in principle, make it easier to attempt integration at the level of the data itself, rather than via Web services, tightly defined query interfaces, or explicit sharing of schemas or models. Nevertheless, successful examples are still lacking. This manuscript describes our attempt to achieve interoperability between two non-coordinating data sources purely through the introduction of shared semantic mapping.

The FAIR principles are a set of guidelines for publication of data and metadata that enhance the ability of datasets to be discovered and processed by machines. The principles focus largely on data reuse, which implicitly requires interoperability.[2] While the FAIR Principles focus largely on metadata, the information that describes the content of data, they also encourage the reformatting of the data into machine-readable syntaxes, with machine-readable semantics (most often implemented using RDF/OWL).

FAIR forms the basis of the EJP-RD Virtual Platform, where several dozen participating rare disease registries and biobanks are being prepared for integrative queries via transformation of their contents into FAIR data formats.[12] As an initial target for interoperability between EJP-RD datasets, 16 Common Data Elements (CDEs) were selected, as defined by the European Platform on Rare Disease Registration.[13] A core FAIR semantic model based on Semantic Science Integrated Ontology (SIO) was designed to act as a scaffold for these Common Data Elements, enabling model reusability and thus simplifying query.[14] All CDE data within the EJP-RD are transformed into RDF that adheres to this core model.

FAIR principles also guide the development of the RDCA-DAP.[15–17] RDCA-DAP provides a neutral environment for industry, academia, regulators, and other government agencies to work together to accelerate and de-risk the medical product development process in rare diseases. RDCA-DAP integrates existing datasets from various sources, including clinical trials, patient registries, preclinical data, natural history studies, and electronic health records. Much of the data in RDCA-DAP originates from clinical trials and is therefore already standardized, using standards from the Clinical Data Interchange Standards Consortium (CDISC). CDISC is a standards development organization that maintains and develops a suite of data standards that encompass data acquisition through analysis.[18] Collected data are represented by the CDISC Study Data Tabulation Model (SDTM), a common data model required for clinical trial data submission to several regulatory agencies.[19] The model provides a standard for the representation of collected data into domains of similar biomedical concepts, such as demographics, laboratory tests, or concomitant medications. Because of its use in regulatory activities, SDTM is strongly aligned with clinical trial (CT) data. The growing importance of real-world data has raised interest in integrating SDTM-formatted CT data with other data types, such as patient registries and electronic health records.[20]

To integrate trial data with other data types and support cross-disease data exploration, an initiative is underway to build semantically grounded models that represent RDCA-DAP data in OWL.

Biolink is a high-level biological domain data model used to represent biomedical entities and the relationship between them. Entities are annotated with ontological terms to semantically ground their meaning.[1] Although Biolink was developed using LinkML (a YAML-based schema modelling language),[21] it has been translated into multiple formats, including an OWL version available through BioPortal.[22] Biolink has links to external ontologies, and it has been used in a variety of health registries and projects, such as the KG-COVID-19 project.[23]

Methods

Coordination was undertaken through weekly meetings by a group with a range of differing expertise, including Semantic Web, Linked Data, and HealthCare data management. The group was formed from representatives of C-Path and EJP-RD organizations, with the larger goal of achieving data interoperability via federation between both organizations' datasets.

For initial analysis, each organization provided an existing dataset about patients with polycystic kidney disease (PKD). The C-Path dataset consists of aggregated data from multiple studies that was gathered by the PKD Outcomes Consortium.[24] These data have been used to develop CDISC data standards for PKD and to support the Food and Drug Administration (FDA) and European Medicines Agency (EMA) qualifications of Total Kidney Volume as an imaging biomarker for drug development tools. The already anonymized dataset was further protected for this study by synthesizing values for laboratory test results using the synthpop R package.[25] The EJP-RD dataset consists of mock data spanning three clinical information domains, described in **Table 1**. These domains were selected from both datasets as the initial targets for federation because they contain semantically similar entities that appear in both the C-Path and EJP-RD semantic data models and datasets. These entities were then mapped to the Biolink model (**Table 2**). National Cancer Institute Thesaurus (NCIT) terms were used for specific fields (e.g., sex, laboratory test names), because study data tabulation model (SDTM) controlled terminologies are already mapped to NCIT, and EJP-RD had already chosen the NCIT OBO version as the reference ontology for several domains.[26]

To execute the study, the shared Biolink concept URLs were added into both C-Path and EJP-RD data models to

Table 1: Selected clinical information types present in both datasets.

Domain	C-Path	EJP-RD
Birthdate/Age patient information	Age as an integer	ISO 8601 compliant date string
Sex patient information	Sex label as a string (F, M) mapped to NCIT terms for Female and Male	NCIT term for Sex (Female, Male, Undetermined, or Unknown) and Sex label as a string
Laboratory data measurements	· Laboratory test name (e.g., Leukocytes), category (e.g., hematology), and specimen type (e.g., blood) · Numerical result and standard ranges in both original and standard units · Study day of lab test · Associated subject visit · Associated specimen collection procedure	· Procedure defined as Quantitation or Estimation · Materials tested input · Target molecular or compound measured · Output measurement value and its unit · ISO 8601 compliant date of measurement procedure

Table 2: Mapping of similar conceptual entities between Biolink, C-Path, and EJP-RD.

Biolink model entities	C-Path SDTM mapping	EJP-RD
Case https://w3id.org/biolink/vocab/Case	Subject https://w3id.org/c-path/biolink_sdtm_owl/SUBJECT	Person http://semanticscience.org/resource/SIO_000498
Procedure https://w3id.org/biolink/vocab/Procedure	Laboratory Test https://w3id.org/c-path/biolink_sdtm_owl/LBTEST and Urinary System Test https://w3id.org/c-path/biolink_sdtm_owl/URTEST	Process http://semanticscience.org/resource/SIO_000006
Information Content Entity https://w3id.org/biolink/vocab/InformationContentEntity	This concept does not exist in the PKD dataset. We reuse https://w3id.org/biolink/vocab/InformationContentEntity	Information Content Entity http://semanticscience.org/resource/SIO_000015
Attribute https://w3id.org/biolink/vocab/Attribute	This concept does not exist in the PKD dataset. We reuse https://w3id.org/biolink/vocab/Attribute	Attribute http://semanticscience.org/resource/SIO_000614
Biological Sex https://w3id.org/biolink/vocab/BiologicalSex	Sex https://w3id.org/c-path/biolink_sdtm_owl/SEX	Sex http://purl.obolibrary.org/obo/NCIT_C28421

harmonize the type of each shared concept, but otherwise the source models were left unchanged. Models for Personal Information and Leukocyte Count are shown for both the EJP-RD (**Figures 1** and **2**) and C-Path (**Figures 3** and **4**). Each project generated RDF data using YARRRML.[27] C-Path ontology, source data, and RDF conversion code are available on GitHub.[28] Both C-Path and EJP-RD data are available on a server.[29]

RDF data from each project were loaded into separate named graphs, such that they could be queried independently. The SPARQL query language was selected as the tool used to explore interoperability between the two datasets. Data can be queried at our SPARQL endpoint by copying or modifying the queries in **Table 3**.[30] A critical feature of SPARQL is that all aspects of a data schema, both entities and relationships, can be represented as variables. This is relevant because the EJP-RD models utilize SIO to describe concept-to-concept relationships, while C-Path uses a customized extension of the Biolink model. Therefore, by leveraging this feature of SPARQL, we attempted to construct near-identical queries over both data models by leaving, as variables, all aspects of the model that are not shared—that is, the queries use only the shared Biolink-typed nodes, leaving all disparate aspects of the underlying models as query variables.

Results

To demonstrate that we have achieved interoperability, SPARQL queries were constructed, with exemplar queries shown in **Table 3**. Query 1 extracts leukocyte count from the C-Path dataset. Query 2 extracts leukocyte count from the EJP-RD dataset. Query 3 uses the SPARQL SERVICE clause to show how a federated query would be constructed over multiple registries; however, in this case, the data is hosted in two separate graphs on the same server.

Discussion

Our results demonstrate that independent mapping of datasets from two distinct data models to an upper-level schema, such as Biolink, offers a starting point of interoperability. In principle, the components of the shared model, such as subjects or patients (e.g., the Case class in Biolink) and their attributes (e.g., the Attribute class in Biolink), provided a detailed-enough structure that independent mapping to these elements provided the means of constructing queries that retrieved records from both datasets using only the shared set of Uniform Resource Identifiers (URI). This is exemplified by the bolded features in Queries 1 and 2 in **Table 3**, showing that we can at least partially anchor the query around the shared typed components while leaving the sub-

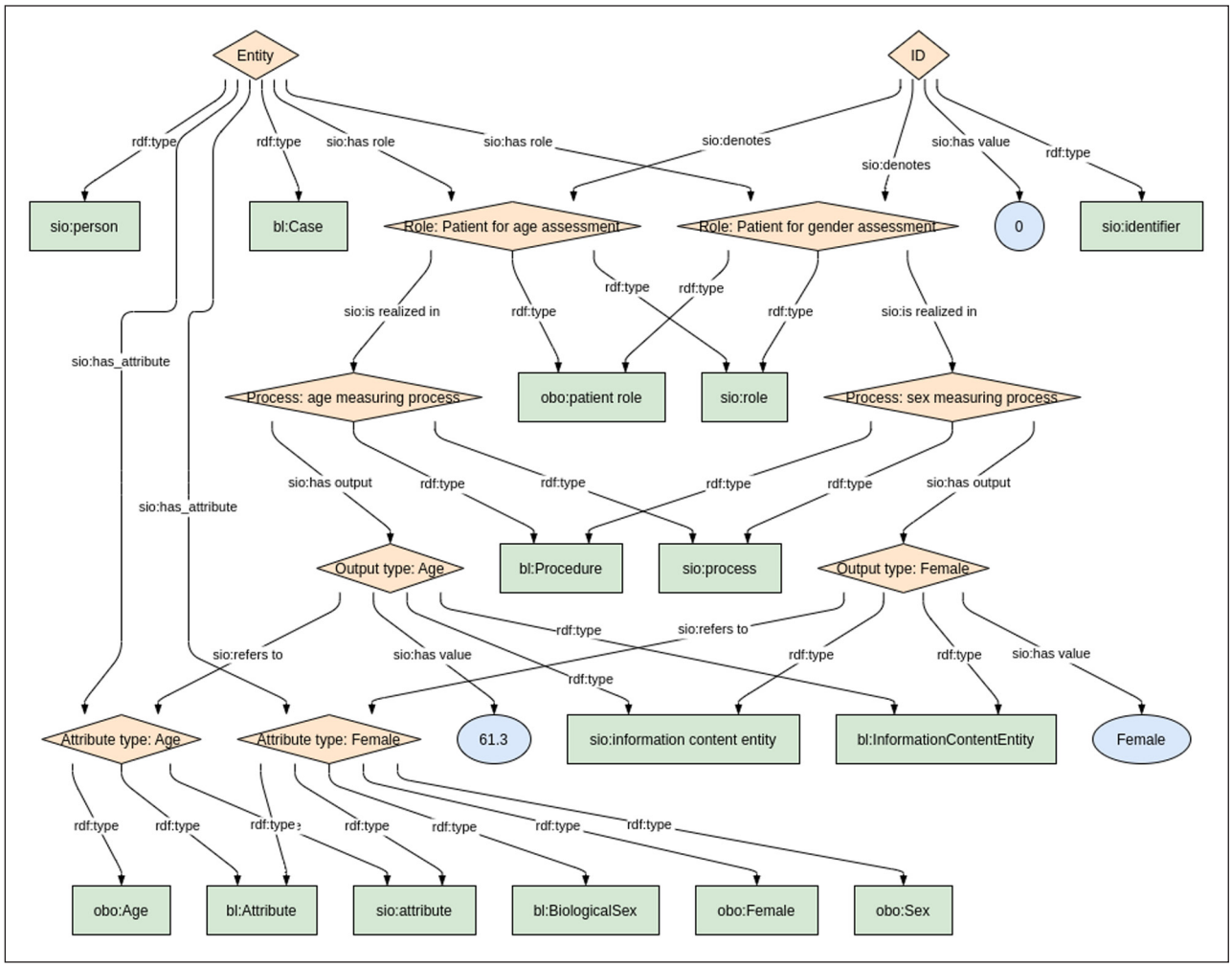


Figure 1: Personal Information model from the EJP-RD Dataset.

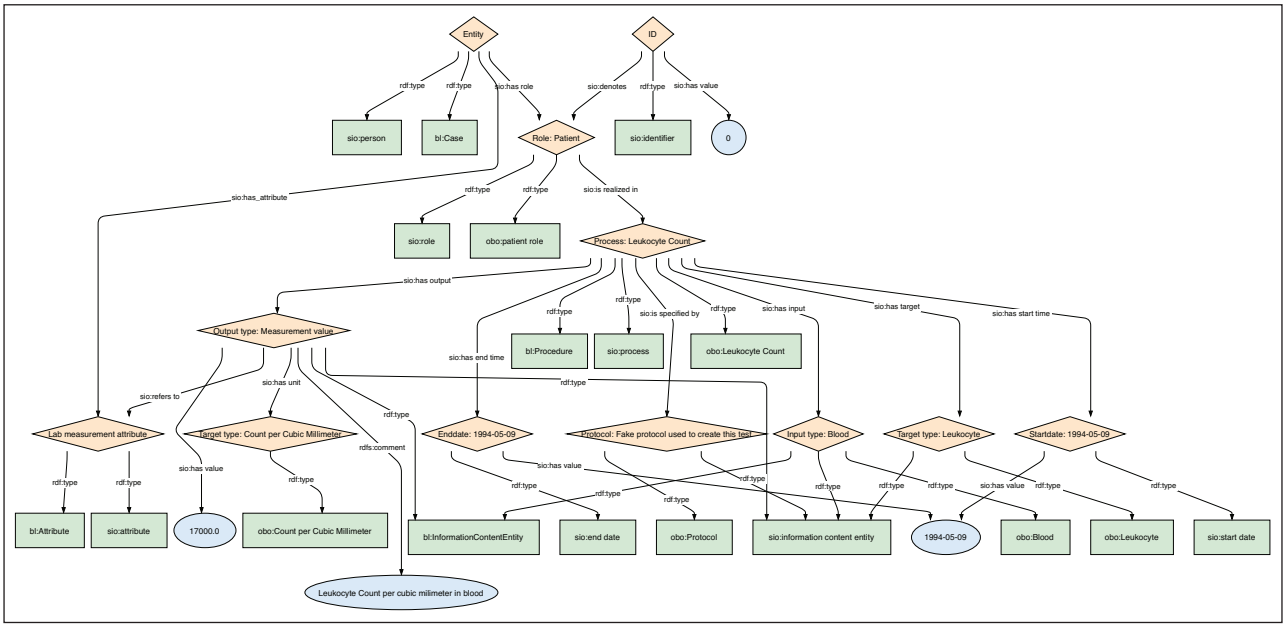


Figure 2: Leukocyte Count Measurement model from the EJP-RD Dataset.

structure of the disparate data models as a query variable. The further extension of this interoperability will require additional collaboration to harmonize how more general shared concepts, such as procedures and activities, relate to the underlying literal values in each dataset.

This early-stage investigation had some limitations. Several actions were taken that, in a realistic scenario, would not be possible. For example, adding an additional type node into the EJP-RD data model was only possible because we are the custodians of that data and, thus, were able to create a

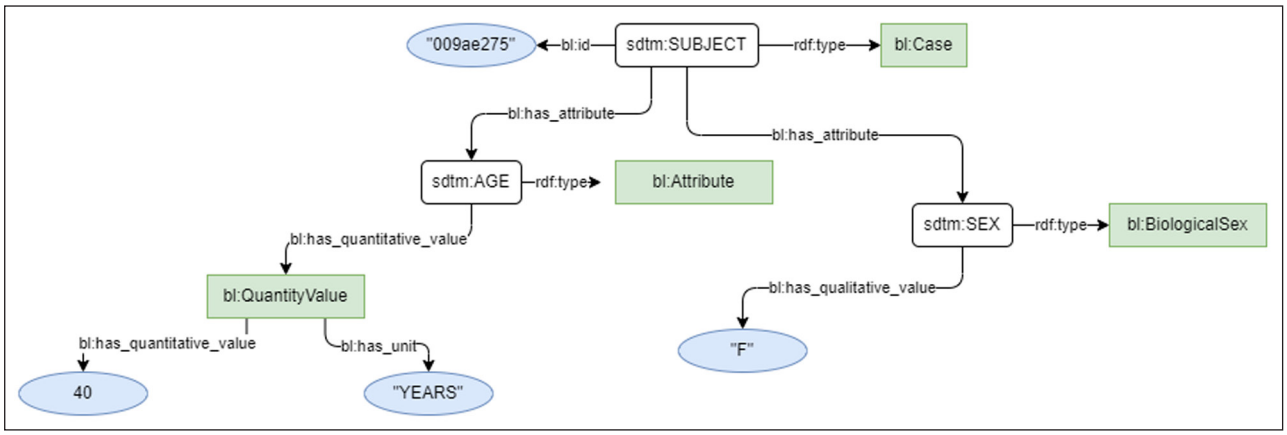


Figure 3: Demographics model for the C-Path SDTM dataset.

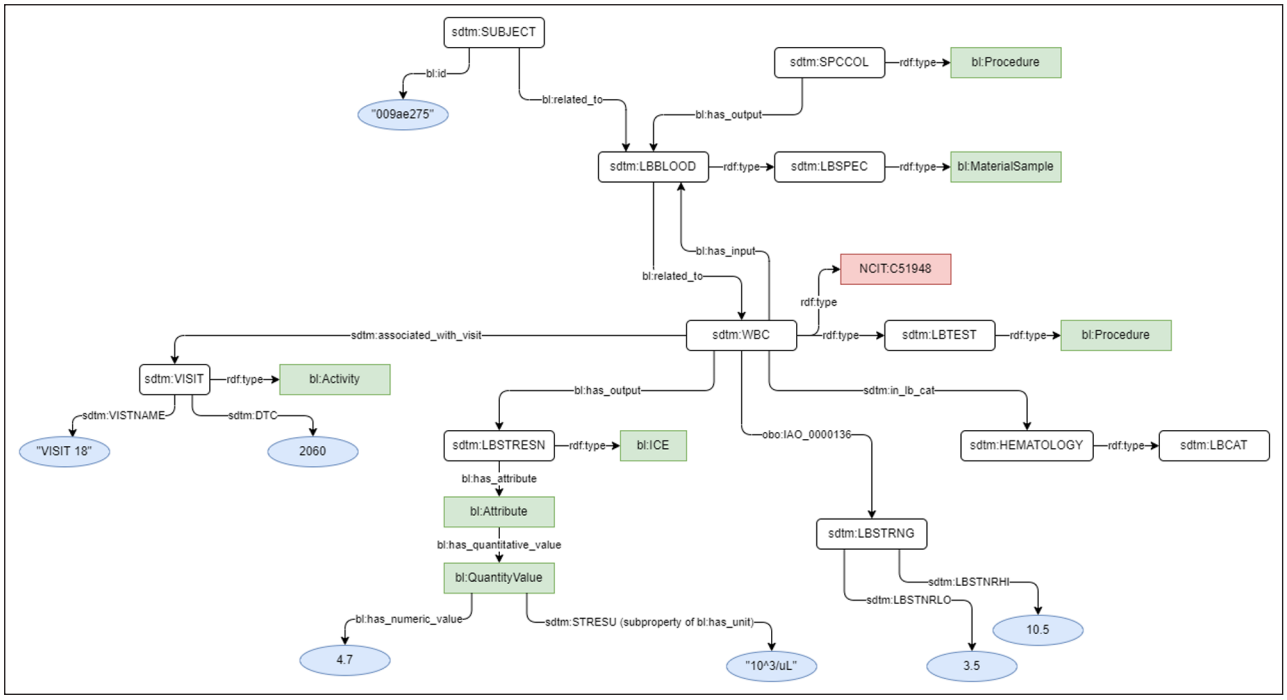


Figure 4: Leukocyte Count Laboratory Test model for the C-Path SDTM dataset.

separate dataset that carried this additional information. This was done to set a baseline for how much interoperability we might expect from this mapping effort. Nevertheless, this direct manipulation of the data is not strictly necessary to achieve the goals pursued here. For example, the mapping in **Table 2** could be captured in RDF and published independently. This would enable a query to be constructed that utilized the mapping data as a third SPARQL SERVICE (i.e., extending Query 3 to make a third SERVICE call to the mapping dataset). This alternative approach would also solve another, more subtle, problem that arises by adding Biolink types directly into the data. That is, Biolink is a schema, not an ontology. In RDF, the type property has a strict semantic meaning and should only be used to categorize a concept based on an ontology term. Thus, by adding Biolink types into the EJP-RD dataset, we render it unusable for the purpose of logical reasoning. Because the C-Path model was built using the OWL version of Biolink, this exception does not apply to the C-Path data, but the limited semantics of Biolink do limit the reasoning that is possible.

Other identified limitations, however, cannot be so easily resolved. One example is the use of age in the personal information from one dataset versus date of birth in the personal information of the other (not shown). While it would be possible to extract both data types using SPARQL (e.g., by simply creating a query that gathers *all* personal information), it is not possible to harmonize the data within the query itself, thus requiring some degree of post-processing of the results. This, however, is not atypical in clinical research or meta-analyses, and thus we still consider the gain in query power of FAIR data to be useful. Other limitations were observed that we will not detail here, such as inconsistent use of units across datasets, data cleaning and reformatting that must be done before conversion to RDF, and the lack of standardized vocabularies for many domains of clinical data. Some can be overcome by enhancing the semantic content within both datasets.

FAIR representations of legacy data require additional curation work, but in our experience, they are generally not more time-consuming than mapping data to less

Table 3: SPARQL Queries demonstrating interoperability (number of records returned in bold beneath each query).

Query 1
Leukocyte Counts from C-Path dataset
	PREFIX ncit: <http://purl.obolibrary.org/obo/>
	PREFIX biol: <http://purl.org/NET/biol/ns#>
	PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
	PREFIX sio: <http://semanticscience.org/resource/>
	PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
	PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
	PREFIX biolink: <https://w3id.org/biolink/vocab/>
	PREFIX bl: <https://w3id.org/biolink/>
	PREFIX blowl: <https://w3id.org/c-path/biolink_sdtm_owl/>
	
	SELECT DISTINCT ?test ?value WHERE {
	GRAPH <http://w3id.org/FAIR_Training_LDP/DAV/home/LDP/cpath/cpath_full> {
	?test a **biolink:Procedure**, ncit:NCIT_C51948 .
	?test **?has_output** ?output .
	?output a **biolink:InformationContentEntity** .
	?output bl:has_attribute ?att .
	?att bl:has_quantitative_value|bl:has_qualitative_value ?valnode .
	?valnode bl:has_numeric_value ?value
	}
	}
	
	514 Records returned from query

Query 2
Leukocyte counts from EJP-RD
	PREFIX ncit: <http://purl.obolibrary.org/obo/>
	PREFIX biol: <http://purl.org/NET/biol/ns#>
	PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
	PREFIX sio: <http://semanticscience.org/resource/>
	PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
	PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
	PREFIX biolink: <https://w3id.org/biolink/vocab/>
	PREFIX bl: <https://w3id.org/biolink/>
	PREFIX blowl: <https://w3id.org/c-path/biolink_sdtm_owl/>
	
	SELECT ?value ?unit WHERE {
	GRAPH <http://w3id.org/FAIR_Training_LDP/DAV/home/LDP/cpath/cbgp_leuk> {
	?test a **biolink:Procedure**, ncit:NCIT_C51948 .
	?test **?has_output** ?output .
	?output a **biolink:InformationContentEntity** .
	?output sio:SIO_000300 ?value .
	?output sio:SIO_000221 ?unitnode .
	?unitnode rdfs:label ?unit
	}
	}
	
	3554 Records returned from query

Query 3
Leukocyte counts from both datasets
	PREFIX ncit: <http://purl.obolibrary.org/obo/>
	PREFIX obo: <http://purl.obolibrary.org/obo/>
	PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
	PREFIX sio: <http://semanticscience.org/resource/>
	PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
	PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
	PREFIX biolink: <https://w3id.org/biolink/vocab/>
	PREFIX bl: <https://w3id.org/biolink/>
	
	SELECT DISTINCT ?test ?value ?unit WHERE {
	{SERVICE <http://fairdata.systems:8890/sparql> {
	{SELECT ?test ?value where {
	GRAPH <http://w3id.org/FAIR_Training_LDP/DAV/home/LDP/cpath/cpath_full> {
	?test a biolink:Procedure, ncit:NCIT_C51948 .
	?test ?has_output ?output .
	?output a biolink:InformationContentEntity .
	?output bl:has_attribute ?att .
	?att bl:has_quantitative_value|bl:has_qualitative_value ?valnode .
	?valnode bl:has_numeric_value ?value
	}
	}
	}
	}
	}

(Contd.)

```
UNION
{SERVICE <http://fairdata.systems:8890/sparql>{
 {SELECT ?test ?value ?unit where {
     GRAPH <http://w3id.org/FAIR_Training_LDP/DAV/home/LDP/cpath/cbgp_leuk> {
          ?test a biolink:Procedure, ncit:NCIT_C51948 .
          ?test ?has_output ?output .
          ?output a biolink:InformationContentEntity .
?output sio:SIO_000300 ?value .
?output sio:SIO_000221 ?unitnode .
          ?unitnode rdfs:label ?unit
   }
  }
 }
}
}
```

4068 Records returned from query

FAIR standards (e.g., that do not use standardized vocabularies or include URLs for their terminology). It is difficult to estimate the total person hours required for us to carry out this project, which took place over about three months, because the most time-consuming parts—building the semantic models and (for C-Path) mapping data to SDTM—were already done. During the three-month project period, our teams met online weekly to make sure we understood each other's models and our mappings to Biolink, because all the models are new and not yet well-known. However, an end goal is to allow independent mappings for improved interoperability.

Conclusions and Future Directions

This is an early report of an initiative to enable interoperability between two large international clinical data sharing initiatives: C-Path and EJP-RD. Biolink was chosen as the enabling technology to establish FAIR Data, and in the limited exploratory datasets created for this study, it was found to enable some degree of cross-compatible federated query over the two datasets. A CDISC contribution that would speed up this and similar work would be the creation of globally unique, permanent, resolvable identifiers for SDTM terms (both standard variables and value terminologies). Such IDs would allow direct reuse of SDTM in knowledge graphs, without having to create duplicate terms in an ontology. The mapping of NCIT terms to CDISC terminology is a step in this direction that allowed us to reuse NCIT for some, and work undertaken in the CDISC 360 initiative is highly relevant in this context.[31] We are examining alternative mechanisms to enhance interoperability while reducing the degree to which any participating dataset must modify its contents, such as publishing an external mapping between the two models (mentioned in the discussion) and better semantic encoding within each dataset (e.g., replacing the sex values in C-Path data with NCIT concepts). We will continue to provide access to semantic models and analysis and data transformation code via public repositories. Data will continue to be available through C-Path's RDCA-DAP and will be delivered via the EJP-RD Virtual Platform when it goes live.

Acknowledgements

PAM, NB, RC, and MDW are supported by funding from the European Union's Horizon 2020 research and innovation programme under the EJP RD COFUND-EJP N° 825575. IB, EH, DO, and RLW are supported by the Critical Path Institute, which is supported by the FDA of the U.S. Department of Health and Human Services (HHS), and is 54.2% funded by the FDA/HHS, totaling $13,239,950, and 45.8% funded by non-government source(s), totaling $11,196,634. The contents are those of the author(s) and do not necessarily represent the official views of, nor are an endorsement by, FDA/HHS or the U.S. Government.

Competing Interests

The authors have no competing interests to declare.

Author Contributions

Pablo Alarcón-Moreno and Ian Braun contributed equally.

References

1. **Biolink Model.** https://biolink.github.io/biolink-model/. Accessed February 2, 2022.
2. **Wilkinson MD,** et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data.* 2016 Mar 15; 3: 160018. Erratum in: *Sci Data.* 2019 Mar 19; 6(1): 6. PMID: 26978244; PMCID: PMC4792175. DOI: https://doi.org/10.1038/sdata.2016.18
3. **Covitz PA, Hartel F, Schaefer C,** et al. caCORE: A common infrastructure for cancer informatics. *Bioinformatics.* 2003; 19(18): 2404–2412. DOI: https://doi.org/10.1093/bioinformatics/btg335
4. **Phillips J, Chilukuri R, Fragoso G, Warzel D, Covitz PA.** The caCORE Software Development Kit: Streamlining construction of interoperable biomedical information services. *BMC Med Inform Decis Mak.* 2006; 6(1): 1–16. DOI: https://doi.org/10.1186/1472-6947-6-2
5. TAPIR–TDWG Access Protocol for Information Retrieval. http://tdwg.github.io/tapir/docs/tdwg_tapir_specification_2010-05-05.html. Accessed January 28, 2022.

6. **Stevens RD, Robinson AJ, Goble CA.** myGrid: Personalised bioinformatics on the information grid. *Bioinformatics*. 2003; 19(suppl_1): i302–i304. DOI: https://doi.org/10.1093/bioinformatics/btg1041
7. **Wilkinson MD, Links M.** BioMOBY: An open source biological web services proposal. *Brief Bioinform*. 2002; 3(4): 331–341. DOI: https://doi.org/10.1093/bib/3.4.331
8. **Wilkinson MD, Senger M, Kawas E,** et al. Interoperability with Moby 1.0—It's better than sharing your toothbrush! *Brief Bioinform*. 2008; 9(3): 220–231. DOI: https://doi.org/10.1093/bib/bbn003
9. **Wilkinson MD, Vandervalk B, Mccarthy L, Wilkinson M.** The Semantic Automated Discovery and Integration (SADI) Web service Design-Pattern, API and Reference Implementation. *Nat Preced 2011*. October 2011; 1–1. DOI: https://doi.org/10.1038/npre.2011.6550.1
10. **Gessler DDG, Schiltz GS, May GD,** et al. SSWAP: A simple semantic web architecture and protocol for semantic web services. *BMC Bioinformatics*. 2009; 10(1): 309. DOI: https://doi.org/10.1186/1471-2105-10-309
11. **Hellerstein JM, Stonebraker M.** Readings in database systems. 2005:865. https://mitpress.mit.edu/books/readings-database-systems-fourth-edition. Accessed February 2, 2022.
12. **European Joint Programme on Rare Diseases.** 2022. What is the Virtual Platform. https://www.ejprarediseases.org/what-is-it/. Accessed April 20, 2022.
13. **Set of common data elements for rare diseases registration.** https://eu-rd-platform.jrc.ec.europa.eu/sites/default/files/CDS/EU_RD_Platform_CDS_Final.pdf. Accessed June 30, 2021.
14. **Dumontier M, Baker CJO, Baran J,** et al. The semanticscience integrated ontology (SIO) for biomedical research and knowledge discovery. *J Biomed Semantics*. 2014; 5(1): 1–11. DOI: https://doi.org/10.1186/2041-1480-5-14
15. **C-Path RDCA-DAP Portal.** https://portal.rdca.c-path.org/. Accessed February 16, 2022.
16. Critical Path Institute. https://c-path.org/. Accessed December 15, 2021.
17. **Woosley RL, Myers RT, Goodsaid F.** The Critical Path Institute's Approach to Precompetitive Sharing and Advancing Regulatory Science. *Clin Pharmacol Ther*. 2010; 87(5): 530–533. DOI: https://doi.org/10.1038/clpt.2010.27
18. **Cdisc.org.** 2022. *Global Regulatory Requirements | CDISC*. https://www.cdisc.org/resources/global-regulatory-requirements. Accessed April 18, 2022.
19. **Cdisc.org.** 2022. *SDTM | CDISC*. https://www.cdisc.org/standards/foundational/sdtm. Accessed April 18, 2022.
20. **Arlett P, Kjær J, Broich K, Cooke E.** Real-World Evidence in EU Medicines Regulation: Enabling Use and Establishing Value. *Clin Pharmacol Ther*. 2022; 111(1): 21–23. DOI: https://doi.org/10.1002/cpt.2479
21. **YAML Ain't Markup Language (YAMLTM) revision 1.2.2.** https://yaml.org/spec/1.2.2/. Accessed February 2, 2022.
22. **Noy NF, Shah NH, Whetzel PL,** et al. BioPortal: Ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res*. 2009; 37(suppl_2): W170–W173. DOI: https://doi.org/10.1093/nar/gkp440
23. **Reese JT, Unni D, Callahan TJ,** et al. KG-COVID-19: A Framework to Produce Customized Knowledge Graphs for COVID-19 Response. Patterns. 2021; 2(1). DOI: https://doi.org/10.1016/j.patter.2020.100155
24. **PKD | Critical Path Institute.** https://c-path.org/programs/pkd/. Accessed February 17, 2022.
25. synthpop: Generating Synthetic Versions of Sensitive Microdata for Statistical Disclosure Control. https://CRAN.R-project.org/package=synthpop
26. **NCI Thesaurus OBO Edition.** https://obofoundry.org/ontology/ncit.html. Accessed February 17, 2022.
27. **YARRRML.** https://rml.io/yarrrml/spec/. Accessed February 2, 2022.
28. GitHub – criticalpathinstitute/biolink_sdtm_owl: An ontology for a proof of concept mapping of PKD data in the SDTM format to the Biolink Model. https://github.com/criticalpathinstitute/biolink_sdtm_owl. Accessed February 17, 2022.
29. **WebDAV Repository.** http://fairdata.systems:8890/DAV/home/LDP/cpath/. Accessed February 17, 2022.
30. **Virtuoso SPARQL Query Editor.** http://fairdata.systems:8890/sparql. Accessed February 17, 2022.
31. **CDISC 360.** https://www.cdisc.org/cdisc-360

How to cite this article: Alarcón-Moreno P, Braun I, Hartley E, Olson D, Benis N, Cornet R, Wilkinson MD, Walls RL. Leveraging Biolink as a FAIR "Rosetta Stone" Between Clinical Semantic Models Provides Emergent Interoperability. *Journal of the Society for Clinical Data Management*. 2022; 2(3): 2, pp. 1–8. DOI: https://doi.org/10.47912/jscdm.130

Submitted: 18 February 2022 **Accepted:** 01 November 2022 **Published:** 23 December 2022

Journal of the Society for Clinical Data Management is a peer-reviewed open access journal published by Society for Clinical Data Management.

OPEN ACCESS

A Resource for Guiding Data Stewards to Make European Rare Disease Patient Registries FAIR

PHILIP VAN DAMME
PABLO ALARCÓN MORENO
CÉSAR H. BERNABÉ
ALBERTO CÁMARA BALLESTEROS
CLÉMENCE M. A. LE CORNEC
BRUNA DOS SANTOS VIEIRA
K. JOERI VAN DER VELDE
SHUXIN ZHANG
CLAUDIO CARTA
RONALD CORNET
PETER A.C. 'T HOEN
ANNIKA JACOBSEN
MORRIS A. SWERTZ
MARCO ROOS
NIRUPAMA BENIS

*Author affiliations can be found in the back matter of this article

RESEARCH PAPER

ABSTRACT

Objective: This paper reports on the development of a dynamic data management planning questionnaire to guide data stewards of the European Reference Network (ERN) rare disease patient registries to make their data findable, accessible, interoperable, and reusable (FAIR). As part of this work, the questionnaire was validated through expert review and aligned with existing resources on rare diseases and FAIR data management.

Materials and Methods: The questionnaire was developed for the Data Stewardship Wizard, a tool for data management planning. Knowledge sources on FAIR data, ERN patient registries, and data management were used to compose questions. Ten domain experts validated the questionnaire. The topics in the questionnaire were aligned with existing knowledge bases.

Results: A total of 57 questions were included in the questionnaire. Twenty-three references to the FAIR Cookbook and Research Data Management toolkit for Life Sciences were added. Expert validation provided a total of 166 comments on content, structure, and software-related issues. A public instance of the Data Stewardship Wizard was deployed for use by data stewards of ERN patient registries.

Discussion: The questionnaire addresses issues that ERNs encounter when making their registries FAIR and follows the implementation choices made by the European rare disease community. A challenging task for future research is to extend the questionnaire to other types of registries and to validate with users.

Conclusion: This smart questionnaire is the first model created for the Data Stewardship Wizard that helps ERN patient registries with making their data FAIR. It will assist data stewards in aligning their efforts and providing guidance on FAIR data.

CORRESPONDING AUTHOR:
Philip van Damme
Amsterdam UMC location University of Amsterdam, Department of Medical Informatics, Meibergdreef 9, Amsterdam, NL; Amsterdam Public Health, Digital Health & Methodology, Amsterdam, NL
p.vandamme@amsterdamumc.nl

KEYWORDS:
FAIR data stewardship; data management planning; rare diseases; European Reference Networks; FAIR data; patient registries

TO CITE THIS ARTICLE:
van Damme, P, Moreno, PA, Bernabé, CH, Ballesteros, AC, Le Cornec, CMA, dos Santos Vieira, B, van der Velde, KJ, Zhang, S, Carta, C, Cornet, R, 't Hoen, PAC, Jacobsen, A, Swertz, MA, Roos, M and Benis, N. 2023. A Resource for Guiding Data Stewards to Make European Rare Disease Patient Registries FAIR. *Data Science Journal*, 19: 12, pp. 1–12. DOI: https://doi.org/10.5334/dsj-2023-012

INTRODUCTION

Up to 36 million people are affected by a rare disease in the European Union (EU), which is around 8% of the total EU population at the time of writing (European Commission 2022b). Like rare disease patients, data about rare diseases are often geographically fragmented. To organize the highly specialized care that patients with a rare disease need, the EU has set up European Reference Networks (ERNs) (European Commission 2017). By exchanging knowledge and information among health care providers, these networks aim to improve access to accurate diagnosis, timely treatment, and appropriate care for people living with rare diseases in Europe.

Members of an ERN share expertise on a specific group of diseases (e.g., rare bone or rare kidney diseases). According to the European Medicines Agency, patient registries collect uniform data over time about a population defined by a particular disease, condition, or exposure (European Medicines Agency 2022). A key task of ERNs is setting up and managing patient registries, which are valuable for research, treatment and outcomes monitoring, drug development, and improving quality of care (Boulanger et al. 2020). Standardizing data management practices to allow for data linking and reuse has been known to increase the benefits of rare disease patient registries (Boulanger et al. 2020; Fink et al. 2017; Kodra et al. 2018). As a result, improving the alignment between ERNs is one of the objectives of the European Joint Programme on Rare Diseases (EJP RD), a project with over 130 institutions from 35 countries, including representatives of all 24 ERNs, designed to establish a self-sustaining infrastructure for rare disease research and care (Inserm 2022). The EJP RD has been supporting patient registries, managed by ERNs, in making informed choices about their data management and in harmonizing choices among registries (Dos Santos Vieira, Bernabé, and Zhang et al. 2022).

Wilkinson et al. (2016) introduced the findable, accessible, interoperable, and reusable (FAIR) principles, a set of high-level guidelines for research data management, stating that data should be FAIR for humans and computers. Dos Santos Vieira et al. (2022) provided insight into common challenges ERNs encountered when making their patient registry data FAIR and put forward a list of solutions that may help solve those challenges. To obtain these insights, a team of data stewards specialized in FAIR data have been working closely with ERN patient registries.

Hudson-Vitale and Moulaison-Sandy (2019) reviewed research on data management plans (DMPs) and reported that DMPs support data sharing and reuse; however, DMPs were often found to be static documents, making them less effective. Williams, Bagwell, and Zozus (2017) presented a framework for DMPs that covers topics such as personnel planning, data elements, data models, software, privacy, and data sharing practices. Since their introduction, the FAIR principles have been a staple for leveraging DMPs that should produce FAIR data. A tool for assembling DMPs is the Data Stewardship Wizard (DSW) (Pergl et al. 2019). The DSW uses dynamic questionnaires that provide context-dependent guidance, can generate DMPs from prebuilt templates, and provide metrics for compliance with the FAIR principles. Moreover, the DSW includes an expertcurated knowledge model, which represents a questionnaire, for creating DMPs for life sciences projects. In addition, Mons (2018) published a book titled *Data Stewardship for Open Science* that encourages readers to create their own DMPs using the DSW.

While the default knowledge models of the DSW are successful in helping data stewards create DMPs for projects in the life sciences domain, they fail to cover the domain-specific requirements of ERN rare disease patient registries. In addition, the European rare disease community, represented by the EJP RD, has made implementation choices for the FAIR principles specific to their domain (Dos Santos Vieira, Bernabé & Zhang et al. 2022). These choices should be reflected in DMPs of rare disease patient registries. Furthermore, sustaining human support for ERNs is challenging, as requirements evolve over time. Hence, there is a strong need for a maintainable data management planning tool tailored to ERN rare disease patient registries.

This paper reports on the results obtained from (1) the creation and validation of a data management planning questionnaire to guide ERN patient registries in making their data FAIR and (2) its integration with existing infrastructures around rare diseases and (FAIR) data management.

MATERIALS AND METHODS

We developed a smart questionnaire for the DSW, that is, a questionnaire with mostly closed-ended questions that adapts follow-up questions based on previously given answers. This questionnaire—to guide data stewards of ERN patient registries to make their registry data FAIR—was built in four stages: (1) collect and analyze relevant knowledge sources; (2) construct a hierarchical mind map from which a DSW knowledge model was built; (3) acquire feedback from domain experts on the different topics in the questionnaire and DMP and align the content with existing tools for FAIR data management planning; and (4) set up a public instance of the DSW with the questionnaire preloaded. Figure 1 summarizes these stages.

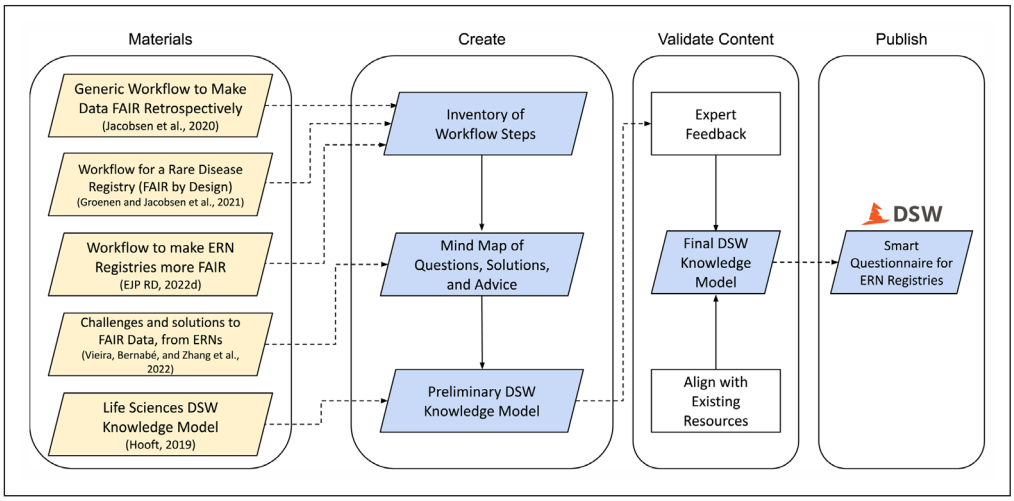


Figure 1 Overview of the stages performed to develop a smart questionnaire for the Data Stewardship Wizard (DSW): gathering relevant knowledge sources, developing a DSW knowledge model (questionnaire), validating the questionnaire, aligning with the Research Data Management toolkit for Life Sciences (RDMkit) (ELIXIR-CONVERGE 2022) and FAIR Cookbook (FAIRplus 2022), and publishing the questionnaire in a DSW instance.

Abbreviations: Data Stewardship Wizard (DSW); European Reference Networks (ERNs); findable, accessible, interoperable, and reusable (FAIR).

Legend: input (yellow), output (blue).

MATERIALS

First, we gathered relevant sources that provide information or knowledge on making ERN patient registries FAIR. These sources guided our subsequent decisions on the topics and questions to be included in the questionnaire.

- Generic workflow (Jacobsen, Kaliyaperumal et al. 2020): a step-by-step workflow to make data that has already been collected FAIR
- Rare disease registry workflow (Groenen and Jacobsen et al. 2021): a workflow designed to make the data of a rare disease patient registry on vascular anomalies FAIR from the moment it is collected
- Workflow to make ERN registries FAIR (EJP RD 2022c): a workflow designed to help ERN patient registries with making their data FAIR
- Challenges and solutions from ERN patient registries (Dos Santos Vieira, Bernabé, and Zhang et al. 2022): an extensive list of 41 challenges and proposed solutions that ERN patient registries encountered when making their data FAIR
- DSW knowledge model for the life sciences domain (Hooft 2019): a knowledge model that includes expert content on data management planning for the life sciences, structured around the research data life cycle

CREATE

We created a preliminary knowledge model in three steps. First, we made an inventory of the steps and implementation choices within the three workflows mentioned under 'Materials'. Second, we built a mind map based on these workflows and ERN challenges. And third, we converted the mind map into a DSW knowledge model. The DSW can export a DMP from a filled-in questionnaire. Questionnaires are generated from knowledge models, which are ordered collections of linked items. A knowledge model contains all the information necessary for generating a questionnaire, such as chapters, questions, descriptions, answer options, and advice bound to answers.

Mind map

We used the workflows for making data FAIR as the basis for a hierarchical mind map. A mind map was considered an appropriate intermediate step before building a knowledge model

because they provide a similar hierarchical structure. This mind map laid the groundwork for what would later become the smart questionnaire. We collaboratively populated the mind map with questions, answer options, and solutions. Solutions were written advice, software tools, standards, references to internal (i.e., EJP RD) and external resources, or other technical solutions for making data of ERN registries FAIR. Questions and solutions were derived from the work of Dos Santos Vieira et al. (2022). We used MindMeister (MeisterLabs 2022), a cloud-based online mind mapping tool.

Knowledge model

After completing the mind map, we converted it to a DSW knowledge model. This step is composed of transferring elements from the mind map and adding additional information (such as answer types, descriptions, and titles) using the DSW's built-in knowledge model editor. In addition, to further enrich our model, we reused all seven chapter names and some relevant questions from the life sciences knowledge model of Hooft (2019). These chapters, based on the research data life cycle (Pergl et al. 2019), were found to be a good fit for structuring the content of our mind map. Hence, our knowledge model was built upon the following chapters: administrative information, reusing data, creating and collecting data, processing data, interpreting data, describing data, and giving access to data. Chapters represent sections of a knowledge model. We restructured questions from the mind map to match the chapters when necessary. Additionally, we added tags to questions that addressed a technical implementation choice for findability, accessibility, interoperability, or reusability. Tags are a feature of the DSW that can be used to organize questions, such as to select questionnaire subsets.

VALIDATE CONTENT

To validate the correctness of the content of the questionnaire, we approached 10 domain experts and asked for their feedback. Among the invited experts were data scientists, project managers, senior researchers, and software engineers. All experts were affiliated with or involved in the EJP RD. Experts had expertise on authentication and authorization, biobanks, data querying, ERNs, FAIR data, patient consent, privacy legislation, project management, rare diseases, rare disease patient registries, record linkage, semantic models, and software architecture. Experts were asked to only appraise content relevant to their expertise. For example, an expert on patient consent would be asked to review all questions related to consent. Experts reviewed individual questions, the structure of the knowledge model, and additional information presented along with the questions and answers. Feedback was collected through a spreadsheet form, via video call, or both. We curated the received expert reviews to remove duplicate comments and to clarify what changes should be made to the knowledge model. We divided the curated feedback into three categories: textual change (question, answer (option), or description), structural change (e.g., change the question order), or software issue. We then updated the knowledge model according to the feedback.

Finally, we aligned the questionnaire with two existing resources that offer a plenitude of knowledge on how to make data FAIR. That is, the Research Data Management toolkit for Life Sciences (RDMkit) and the FAIR Cookbook (ELIXIR-CONVERGE 2022; FAIRplus 2022). We added references to pages from the RDMkit or recipes from the FAIR Cookbook to any description or advice in the questionnaire that mentioned a topic also covered by one or both resources.

PUBLISH

Publishing involved hosting a public instance of the DSW with our knowledge model preloaded. We hosted this instance on the servers of ELIXIR's Czech Republic node, which also manages support and operation of the DSW (Czech National Infrastructure for Biological Data 2022). Existing privacy policies apply to this instance. The knowledge model source files were made available on a public repository (see 'Data Availability' section).

RESULTS

The inventory of workflow steps to make data FAIR and implementation choices suggested by the EJP RD comprises nine steps, 19 topics related to those steps, and 12 implementation choices (e.g., a certain tool or standard). Table 1 shows an overview of this inventory.

WORKFLOW STEP	RELATED TOPICS	IMPLEMENTATION
1. Identify FAIR objectives and expertise	a. Defining objectives	
	b. Giving training	
	c. Hiring of personnel	
2. Define data elements to be collected	a. Common data elements	CDE core elements (European Commission 2019)
	b. Data dictionary	
	c. Central metadata repository registration	ERDRI.mdr (European Commission 2022a)
3. Define metadata elements to be collected	a. Machine interpretable metadata	EJP RD metadata model (EJP RD 2022d)
	b. Metadata store	FAIR data point (Bonino da Silva Santos et al. 2023)
4. Create a semantic data model	a. Reuse of existing model(s)	CDE semantic model (Kaliyaperumal et al. 2022) CDISC ODM (CDISC 2022) HL7 FHIR (HL7 2022) OMOP CDM (OHDSI 2022)
5. Obtain consent	a. Standardized informed consent form	ERN ICF (EJP RD 2022b)
6. Enter (FAIR) data	a. Electronic data capture systems	
7. Standardize metadata	a. Metadata model(s)	EJP RD metadata model (EJP RD 2022d)
	b. Standard terminology	CDE semantic model terminology (EJP RD 2022e)
8. Transform (meta)data to RDF	a. Data transformation	CDE in a box (EJP RD 2022a)
	b. Terminology mappings	
9. Manage authentication and authorization	a. Authorization roles	
	b. Access conditions	
	c. Data pseudonymization	
	d.Querying	

Table 1 Overview of the workflow steps and inventory of topics and implementations.

Abbreviations: common data elements on rare disease registration (CDE); European Platform on Rare Disease Registration metadata repository (ERDRI.mdr); European Reference Network (ERN); Health Level 7 Fast Healthcare Interoperability Resources (HL7 FHIR); Clinical Data Interchange Standards Consortium Operational Data Model (CDISC ODM); Observational Medical Outcomes Partnership Common Data Model (OMOP CDM); findable, accessible, interoperable, and reusable (FAIR); informed consent form (ICF); Resource Description Framework (RDF).

Steps, topics, and implementations were translated into questions, answers, and advice. For example, the topic 'defining objectives' was rephrased as 'Have you defined objectives?' Eventually, the mind mapping process resulted in 22 out of 41 ERN challenges identified by Dos Santos Vieira et al. (2022) being included as a question, answer, advice, or a combination of the three. Challenges that were categorized under 'community,' that is, alignment between ERNs, were not included as questions but were indirectly addressed by using the DSW. That is to say, enabling ERN data stewards to use the DSW with our questionnaire untangles those challenges in part. For example, ERNs found that they were unaware of choices other ERNs made, which they can now share through the DSW. Similarly, one challenge categorized under 'training' was also indirectly addressed (need for more information on activities of the EJP RD).

Table 2 shows the number of challenges included in the questionnaire and the motivation for why some were not. The full list of challenges and can be found in the original publication (Dos Santos Vieira, Bernabé, and Zhang et al. 2022).

The mind map was converted into a preliminary DSW knowledge model. That is, questions, answers, and advice were added to the knowledge model based on the mind map. We reused one question from the life sciences knowledge model of Hooft (2019): 'Who is a contributor to the DMP?' Once this preliminary version of the questionnaire was available in the DSW, we started the validation process.

Validating the correctness of the content of the questionnaire resulted in an updated version of our knowledge model. A total of 10 experts reviewed the content of the questionnaire. We received a total of 166 comments. Each chapter was assigned at least seven experts. All experts reviewed the questions in 'Processing data' and 'Interpreting data.' Table 3 shows an overview

CATEGORY	DIRECTLY INCLUDED	INDIRECTLY INCLUDED	MOTIVATION
Community	0 out of 7	7 out of 7	All challenges addressed a lack of alignment between registries. The DSW questionnaire solves this issue.
Implementation	7 out of 9	0 out of 9	Two not-included challenges were irrelevant at the time of developing the questionnaire.
Legal	3 out of 5	0 out of 5	Two not-included challenges addressed a tool that was not relevant for developing the questionnaire.
Modeling	3 out of 5	0 out of 5	Two not-included challenges addressed issues that were too specific.
			Five not-included challenges addressed irrelevant tools.
Training	9 out of 15	1 out of 15	One indirectly covered challenge was not mentioned specifically in the questionnaire but could be deducted from the information.
All categories	22 out of 41	8 out of 41	

Table 2 Challenges and categories from Dos Santos Vieira et al. (2022) that were included in the questionnaire during the mind mapping phase. Challenges marked as indirectly covered are not specifically mentioned in the questionnaire but were solved solely by the use of the Data Stewardship Wizard (DSW) and the questionnaire.

CHAPTER	TEXTUAL CHANGES	STRUCTURAL CHANGES	SOFTWARE ISSUES
Administrative information	18	5	3
Reusing data	13	2	0
Creating and collecting data	3	2	2
Processing data	19	3	0
Interpreting data	47	8	0
Describing data	10	1	0
Giving access to data	27	3	0
All chapters	137	24	5

Table 3 Quantification of the received feedback per chapter. Feedback is categorized as textual change, structural change, or software issue.

of the comments per chapter and category. Duplicate comments often regarded textual issues on flow or clarity. Experts also provided references to additional resources, such as web pages with more information on a certain topic. Structural changes asked for moving a question up or down the hierarchy or to another chapter. Software issues were related to issues with using the DSW interface, such as a nonfunctional button or a page that would not load. These issues were solved by updating to the latest version of the DSW.

After processing the feedback and updating the knowledge model, the questionnaire has 57 questions. A total of 6 questions are open-ended, and 51 questions are closed-ended. In total, 10 references were added to recipes in the FAIR Cookbook and 13 references to pages of the RDMkit. Three questions were tagged as an implementation choice for findability, 6 to accessibility, 14 to interoperability, and 21 to reusability. Thirteen questions were not tagged because they did not cover implementation choices but rather aspects like training, objectives, or administrative topics. Table 4 shows the number of questions and external references per chapter.

CHAPTER	TOP-LEVEL QUESTIONS	TOTAL QUESTIONS	REFERENCES TO FAIR COOKBOOK	REFERENCES TO RDMKIT
Administrative information	6	15	1	4
Reusing data	2	9	3	3
Creating and collecting data	2	5	1	1
Processing data	1	5	0	2
Interpreting data	2	12	4	1
Describing data	2	4	0	0
Giving access to data	4	7	1	2
All chapters	19	57	10	13

Table 4 Questions and external references per chapter. Top-level questions are questions that precede all other questions and are always presented to a user.

Abbreviations: findable, accessible, interoperable, and reusable (FAIR); Research Data Management toolkit for Life Sciences (RDMkit) (ELIXIR-CONVERGE 2022); FAIR Cookbook (FAIRplus 2022).

Figure 2 depicts a simplified view of our knowledge model and includes all topics covered by the questionnaire. This is the final model that was constructed after expert validation. The questionnaire covers a broad range of topics: building and training a team of professionals, defining data management objectives, (meta)data modeling, data elements, using common standards, using common terminology, data pseudonymization, electronic data capture, querying, metadata exposure, authentication and authorization, and informed consent. Figure 3 provides a screenshot of the chapters and top-level questions. Figure 4 provides a screenshot of how the questionnaire is presented to a user.

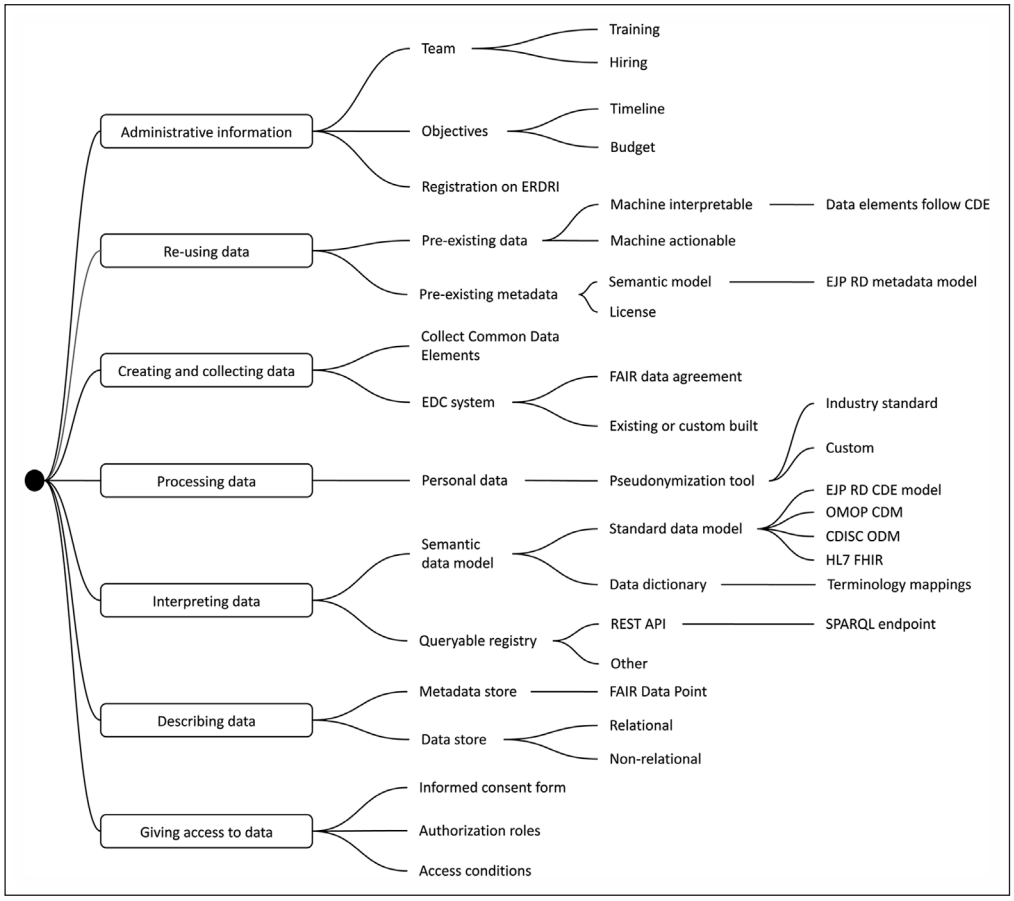


Figure 2 Simplified view of the knowledge model.

Abbreviations: common data elements (CDE); electronic data capture (EDC); European Joint Programme on Rare Diseases (EJP RD); European Platform on Rare Disease Registration (ERDRI); findable, accessible, interoperable, and reusable (FAIR); Health Level 7 Fast Healthcare Interoperability Resources (HL7 FHIR); Clinical Data Interchange Standards Consortium Operational Data Model (CDISC ODM); Observational Medical Outcomes Partnership Common Data Model (OMOP CDM); REpresentational State Transfer Application Programming Interface (REST API); SPARQL Protocol and RDF Query Language (SPARQL).

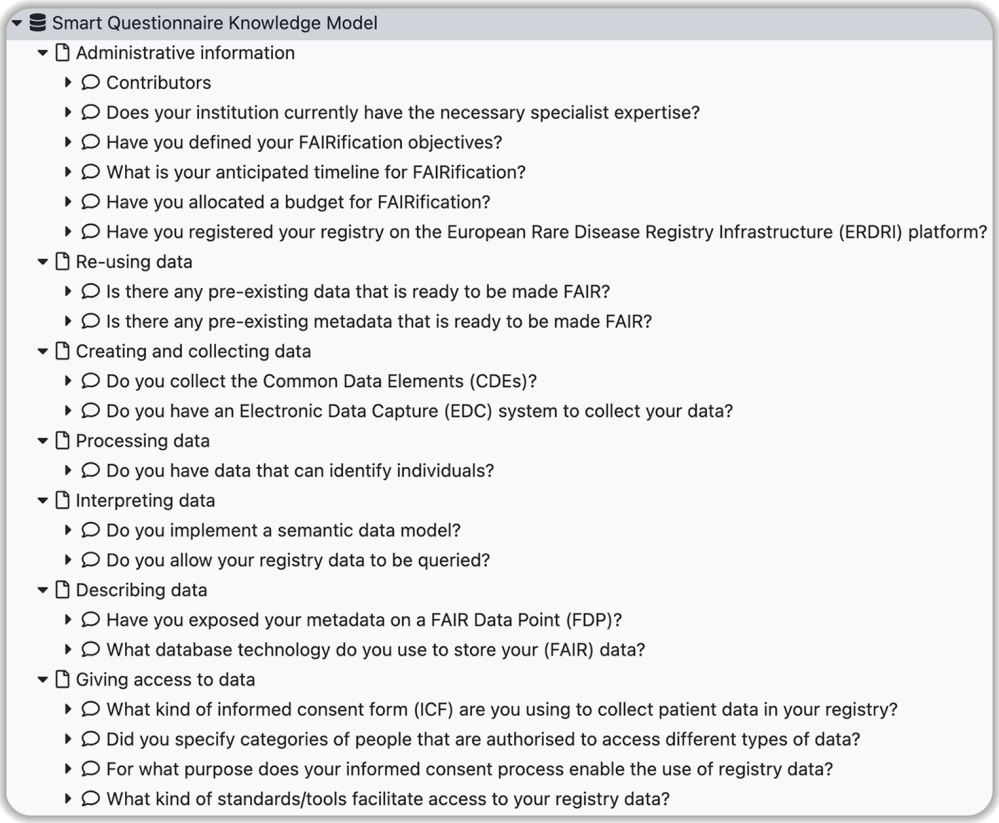


Figure 3 Screenshot of the knowledge model with top-level questions (Data Stewardship Wizard knowledge model editor module).

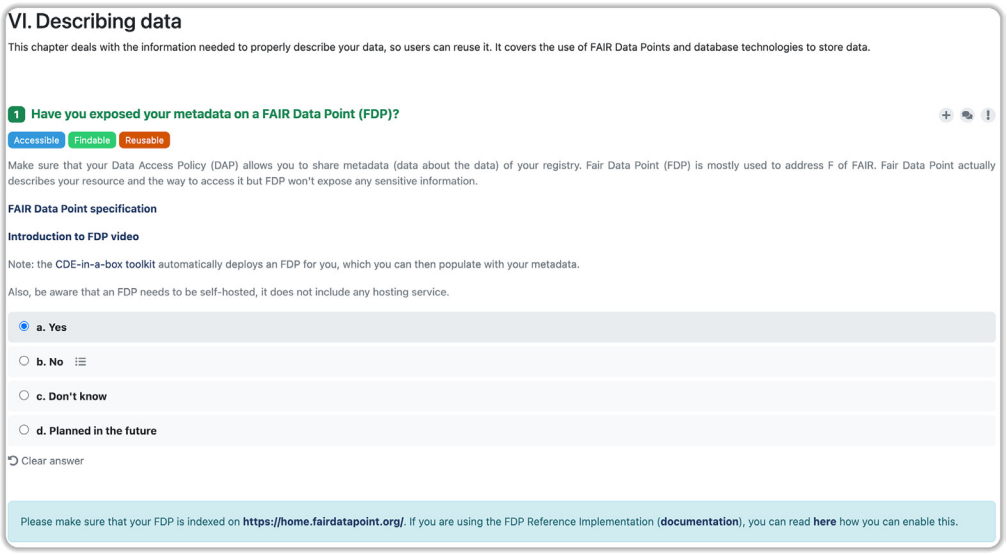


Figure 4 Screenshot of the first question of the 'Describing data' chapter (Data Stewardship Wizard questionnaire module).

DISCUSSION

The purpose of this work was to develop a smart questionnaire that guides data stewards working to make data of ERN rare disease patient registries FAIR. Data stewards of patient registries will increasingly have to manage data in ways that comply with implementation choices of the FAIR principles as recommended by their community. Standardizing data management practices of patient registries enables virtual pooling of otherwise sparse and geographically scattered rare disease data, increasing their usefulness for effective research and care. However, standardization for each of the FAIR principles in this domain is complex. ERNs face the challenge of registering data of thousands of diseases from many different sources and making that data as usable as possible within a global health data ecosystem. Our questionnaire addresses those challenges ERNs were known to commonly face and provides guidance according to the FAIR infrastructure set up by the EJP RD. The questionnaire acts as a checklist for making rare disease registries FAIR: data stewards can make sure that all boxes are checked.

Our questionnaire covers the process of making data in patient registries FAIR. Although the questions and advice are made for ERN patient registries, the questionnaire content was based on prior knowledge of FAIR data and experiences in the European rare disease community. For instance, annual Bring Your Own Data workshops have brought together FAIR data experts and rare disease data managers since 2014 (Roos et al. 2014). Workshops such as these have provided a valuable source of challenges regarding the FAIR guiding principles and proposed implementations thereof (e.g., Jacobsen, de Miranda Azevedo et al. 2020a). Furthermore, they affirmed our motivation for designing a smart questionnaire that enables data stewards to begin their FAIR journey from a variety of starting points. As a result of tailoring each topic and question to the unique needs of ERN patient registries, we have filled the gap in having a data management tool that is suitable for rare disease registries in Europe. Since this work is part of ongoing efforts of the EJP RD, integration of the DSW and questionnaire with the European infrastructure for rare disease research will be a natural next step.

Previous studies focused on DMP requirements (Williams, Bagwell & Zozus 2017) and DMPs for the life sciences domain (Pergl et al. 2019; Hooft 2019). Williams, Bagwell, and Zozus (2017) concluded that while most DMPs included components describing data reuse and sharing, few DMPs described data collection and processing practices. These last two are particularly hard to fix, as the quality of poorly collected data can most likely not be improved in retrospect. We were able to address all four topics in our questionnaire. Creating DMPs for projects in the life sciences was addressed by the original authors of the DSW. We found that by reusing parts of their knowledge model, we were able to structure our questionnaire according to a well-established model. Moreover, Jones et al. (2020) concluded that DMPs are essential for FAIR data stewardship. By adopting the DSW as a tool for making ERN patient registries FAIR, we believe our work aligns with that conclusion.

Our work has some limitations. We validated the content of the questionnaire for correctness through expert feedback, but we did not validate the impact of the questionnaire on its intended users. Therefore, further research is needed to determine whether ERN registry data stewards benefit from our tool. Furthermore, the questions and advice are specific to the situation of ERN patient registries and cannot be extrapolated to other registries or projects without modifications. Our work mainly focused on guiding ERN patient registries in making their data FAIR; nevertheless, there is clear value in aligning more types of registries as well. Registry types outside of rare diseases, as well as non-European rare disease patient registries, could fall into this category.

Our work also has several strengths. First, navigating through FAIR implementation choices via questions and answers is a different experience from filling out DMP checklists. It is anticipated that this will lead to an increase in the quality of DMPs. Second, through the DSW, data stewards can learn from the implementation choices of others. Thus, it complements in-person training and contributes to community convergence. Third, we created a single place where ERN data stewards can go for guidance on making their registry FAIR. This makes maintaining and updating the knowledge in the questionnaire easier compared to having various sources in different locations. The knowledge model can be improved by learning from users who will fill out questionnaires on the DSW platform. Moreover, the DSW software is being actively maintained, and hosting our instance on the ELIXIR infrastructure means that it can be sustained beyond the lifetime of the EJP RD.

The knowledge model we developed is publicly available (see 'Data Availability' section) and can be used by others to build upon or to reuse parts from. For exporting the DMP, we use a default DSW template, which we intend to customize in the near future. It may be possible to improve the guidance offered to ERN data stewards through further customization of this template. Additional research is needed to quantify the impact of our questionnaire on the (process leading to) 'FAIRness' of ERN patient registries. Another challenging task for further research is to extend the questionnaire to other types of resources by collaborating with resource owners and users.

CONCLUSIONS

The developed smart questionnaire for the DSW is a promising method for guiding data stewards in making their registry data FAIR. It is the first model created for the DSW that helps to standardize data management practices among ERN patient registries. Future research should focus on user validation and extending the questionnaire beyond the realm of ERNs.

DATA ACCESSIBILITY STATEMENT

The smart questionnaire is available at https://smartguidance.ejprarediseases.org. (Registration is required.) A user guide and the knowledge model source files are available at https://github.com/ejp-rd-vp/smart-guidance.

ACKNOWLEDGEMENTS

We thank Marek Suchánek and the ELIXIR Czech Republic node for providing technical support and resources for hosting our public instance of the DSW. We thank Rajaram Kaliyaperumal for setting up and maintaining a local instance of the DSW for testing and development and for providing technical support. We thank Anthony Brookes, Esther van Enckevort, Tala Haddad, Rajaram Kaliyaperumal, Karl Kreiner, Nawel Lalout, Annalisa Landi, Yanis Mimouni, and David Reinert for providing feedback on the questionnaire.

FUNDING INFORMATION

This work was supported by the EU's Horizon 2020 research and innovation program under the EJP RD COFUND-EJP No., 825575.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Philip van Damme: conceptualization, methodology, visualization, writing—original draft. Pablo Alarcón Moreno: conceptualization, methodology, writing—review and editing. César H. Bernabé: conceptualization, methodology, writing—review and editing. Alberto Cámara Ballesteros: conceptualization, methodology, writing—review and editing. Clémence M. A. Le Cornec: methodology, writing—review and editing. Bruna Dos Santos Vieira: conceptualization, methodology, writing—review and editing. Joeri van der Velde: methodology, writing—review and editing. Shuxin Zhang: conceptualization, methodology, writing—review and editing. Claudio Carta: writing—review and editing. Ronald Cornet: writing—review and editing. Peter A. C. 't Hoen: writing—review and editing. Annika Jacobsen: resources, writing—review and editing. Morris A. Swertz: writing—review and editing. Marco Roos: conceptualization, project administration, writing—review and editing. Nirupama Benis: conceptualization, methodology, supervision, writing—original draft

AUTHOR AFFILIATIONS

Philip van Damme orcid.org/0000-0002-7124-8949
Amsterdam UMC location University of Amsterdam, Department of Medical Informatics, Meibergdreef 9, Amsterdam, NL; Amsterdam Public Health, Digital Health & Methodology, Amsterdam, NL

Pablo Alarcón Moreno orcid.org/0000-0001-5974-589X
Departamento de Biotecnología-Biología Vegetal, Escuela Técnica Superior de Ingeniería Agronómica, Alimentaria y de Biosistemas, Centro de Biotecnología y Genómica de Plantas. Universidad Politécnica de Madrid (UPM)–Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria-CSIC (INIA-CSIC), Campus Montegancedo 28223 Pozuelo de Alarcón (Madrid), Spain, ES

César H. Bernabé orcid.org/0000-0003-1795-5930
Department of Human Genetics, Leiden University Medical Center, Leiden, NL

Alberto Cámara Ballesteros orcid.org/0000-0001-5613-9704
Departamento de Biotecnología-Biología Vegetal, Escuela Técnica Superior de Ingeniería Agronómica, Alimentaria y de Biosistemas, Centro de Biotecnología y Genómica de Plantas. Universidad Politécnica de Madrid (UPM)–Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria-CSIC (INIA-CSIC), Campus Montegancedo 28223 Pozuelo de Alarcón (Madrid), Spain, ES

Clémence M. A. Le Cornec
Division of Paediatric Nephrology, Centre for Paediatrics and Adolescent Medicine, University of Heidelberg, Heidelberg, Germany, DE

Bruna Dos Santos Vieira orcid.org/0000-0001-7893-0505
Center for Molecular and Biomolecular Informatics, Radboud University Medical Center, Nijmegen, NL; Department of Medical Imaging, Radboud University Medical Center, Nijmegen, NL

K. Joeri van der Velde orcid.org/0000-0002-0934-8375
Genomics Coordination Center, University of Groningen and University Medical Center, Groningen, NL

Shuxin Zhang orcid.org/0000-0003-4715-9070
Amsterdam UMC location University of Amsterdam, Department of Medical Informatics, Meibergdreef 9, Amsterdam, NL; Amsterdam Public Health, Digital Health & Methodology, Amsterdam, NL

Claudio Carta orcid.org/0000-0003-3545-198X
Instituto Superiore di Sanità, Rome, IT

Ronald Cornet orcid.org/0000-0002-1704-5980
Amsterdam UMC location University of Amsterdam, Department of Medical Informatics, Meibergdreef 9, Amsterdam, NL; Amsterdam Public Health, Digital Health & Methodology, Amsterdam, NL

Peter A.C. 't Hoen orcid.org/0000-0003-4450-3112
Center for Molecular and Biomolecular Informatics, Radboud University Medical Center, Nijmegen, NL

Annika Jacobsen orcid.org/0000-0003-4818-2360
Department of Human Genetics, Leiden University Medical Center, Leiden, NL

Morris A. Swertz orcid.org/0000-0002-0979-3401
Genomics Coordination Center, University of Groningen and University Medical Center, Groningen, NL

Marco Roos orcid.org/0000-0002-8691-772X
Department of Human Genetics, Leiden University Medical Center, Leiden, NL

Nirupama Benis orcid.org/0000-0002-2101-6154
Amsterdam UMC location University of Amsterdam, Department of Medical Informatics, Meibergdreef 9, Amsterdam, NL; Amsterdam Public Health, Digital Health & Methodology, Amsterdam, NL

REFERENCES

Bonino da Silva Santos, LO, Burger, K, Kaliyaperumal, R and **Wilkinson, MD.** 2023. FAIR data point: a FAIR-oriented approach for metadata publication. *Data Intelligence*, 5(1): 163–183. DOI: https://doi.org/10.1162/dint_a_00160

Boulanger, V, Schlemmer, M, Rossov, S, Seebald, A and **Gavin, P.** 2020. Establishing patient registries for rare diseases: rationale and challenges. *Pharmaceutical Medicine*, 34(3): 185–190. DOI: https://doi.org/10.1007/s40290-020-00332-1

CDISC. 2022. *Clinical Data Interchange Standards Consortium Operational Data Model*. https://www.cdisc.org/standards/data-exchange/odm

Czech National Infrastructure for Biological Data 2022. *ELIXIR CZ*. https://www.elixir-czech.cz/

Dos Santos Vieira, B, Bernabé, CH, Zhang, S, et al. 2022. Towards FAIRification of sensitive and fragmented rare disease patient data: challenges and solutions in European Reference Network registries. *Orhpanet Journal of Rare Diseases*. DOI: https://doi.org/10.21203/rs.3.rs-1572508/v1

EJP RD. 2022a. *CDE-in-box*. https://github.com/ejp-rd-vp/cde-in-box

EJP RD. 2022b. *Semantic data model of the set of common data elements for rare disease registration*. https://github.com/ejp-rd-vp/CDE-semantic-model

EJP RD. 2022c. *Metadata for EJP rare disease patient registries, biobanks and catalogs*. https://github.com/ejp-rd-vp/resource-metadata-schema

EJP RD. 2022d. FAIRopoly—*FAIRification guidance for ERN patient registries*. https://www.ejprarediseases.org/fairopoly/

EJP RD. 2022e. *ERN registries generic informed consent forms*. https://www.ejprarediseases.org/ern-registries-generic-icf/

ELIXIR-CONVERGE. 2022. *The Research Data Management toolkit for Life Sciences (RDMkit)*. https://rdmkit.elixir-europe.org/

European Commission. 2017. *European Reference Networks: working for patients with rare, low-prevalence and complex diseases*. https://health.ec.europa.eu/publications/brochure-european-reference-networks-rare-and-complex-diseases_en

European Commission. 2019. *Set of common data elements*. https://eu-rd-platform.jrc.ec.europa.eu/set-of-common-data-elements_en

European Commission. 2022a. *Rare diseases*. https://health.ec.europa.eu/non-communicable-diseases/steering-group/rare-diseases_en

European Commission. 2022b. *European Platform on Rare Disease Registration metadata repository (ERDRI.mdr)*. https://eu-rd-platform.jrc.ec.europa.eu/mdr/

European Medicines Agency. 2022. *Patient registries*. https://www.ema.europa.eu/en/human-regulatory/post-authorisation/patient-registries

FAIRplus. 2022. *The FAIR Cookbook*. https://faircookbook.elixir-europe.org/

Fink, AK, Loeffler, DR, Marshall, BC, Goss, CH and **Morgan, WJ.** 2017. Data that empower: the success and promise of CF patient registries. *Pediatric Pulmonology*, 52: S44–S51. DOI: https://doi.org/10.1002/ppul.23790

Groenen, KHJ, Jacobsen, A, Kersloot, MG, et al. 2021 The de novo FAIRification process of a registry for vascular anomalies. *Orphanet Journal of Rare Diseases*, 16. DOI: https://doi.org/10.1186/s13023-021-02004-y

HL7. 2022 Health Level 7 Fast Healthcare Interoperability Resources. https://www.hl7.org/fhir/

Hooft, RWW. 2019. Data stewardship mindmap. https://doi.org/10.5281/zenodo.2614819

Hudson-Vitale, C and **Moulaison-Sandy, H.** 2019. Data management plans: a review. *DESIDOC Journal of Library and Information Technology*, 39(6): 322–328. DOI: https://doi.org/10.14429/djlit.39.06.15086

Inserm. 2022. *European Joint Programme on Rare Diseases*. https://www.ejprarediseases.org/

Jacobsen, A, de Miranda Azevedo, R, Juty, N, et al. 2020. FAIR principles: interpretations and implementation considerations. *Data Intelligence*, 2(1–2): 10–29. DOI: https://doi.org/10.1162/dint_r_00024

Jacobsen, A, Kaliyaperumal, R, Bonino da Silva Santos, LO, Mons, B, Schultes, E, Roos, M and **Thompson, M.** 2020. A generic workflow for the data FAIRification process. *Data Intelligence*, 2(1–2): 56–65. DOI: https://doi.org/10.1162/dint_a_00028

Jones, S, Pergl, R, Hooft, R, et al. 2020. Data management planning: how requirements and solutions are beginning to converge. *Data Intelligence*, 2(1–2): 208–219. DOI: https://doi.org/10.1162/dint_a_00043

Kaliyaperumal, R, Wilkinson, MD, Moreno, PA, et al. 2022. Semantic modelling of common data elements for rare disease registries, and a prototype workflow for their deployment over registry data. *Journal of Biomedical Semantics*, 13. DOI: https://doi.org/10.1186/s13326-022-00264-6

Kodra, Y, Weinbach, J, Posada-de-la-Paz, M, et al. 2018. Recommendations for improving the quality of rare disease registries. *International Journal of Environmental Research and Public Health*, 15(8): 1644. DOI: https://doi.org/10.3390/ijerph15081644

MeisterLabs. 2022. *MindMeister*. https://www.mindmeister.com/

Mons, B. 2018. *Data Stewardship for Open Science*. New York: Chapman and Hall/CRC. DOI: https://doi.org/10.1201/9781315380711

OHDSI. 2022. *Observational Medical Outcomes Partnership Common Data Model*. https://www.ohdsi.org/data-standardization/

Pergl, R, Hooft, R, Suchánek, M, Knaisl, V and **Slifka, J.** 2019. 'Data Stewardship Wizard': a tool bringing together researchers, data stewards, and data experts around data management planning. *Data Science Journal*, 18(1): 59. DOI: https://doi.org/10.5334/dsj-2019-059

Roos, M, Gray, AJG, Waagmeester, A, et al. 2014. Bring Your Own Data workshops: a mechanism to aid data owners to comply with Linked Data best practices. https://ceur-ws.org/Vol-1320/paper_36.pdf

Wilkinson, MD, Dumontier, M, Aalbersberg, IJ, et al. 2016. Comment: the FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3. DOI: https://doi.org/10.1038/sdata.2016.18

Williams, M, Bagwell, J and **Zozus, MN.** 2017. Data management plans, the missing perspective. *Journal of Biomedical Informatics*, 71: 130–142. DOI: https://doi.org/10.1016/j.jbi.2017.05.004

TO CITE THIS ARTICLE:
van Damme, P, Moreno, PA, Bernabé, CH, Ballesteros, AC, Le Cornec, CMA, dos Santos Vieira, B, van der Velde, KJ, Zhang, S, Carta, C, Cornet, R, 't Hoen, PAC, Jacobsen, A, Swertz, MA, Roos, M and Benis, N. 2023. A Resource for Guiding Data Stewards to Make European Rare Disease Patient Registries FAIR. *Data Science Journal*, 19: 12, pp. 1–12. DOI: https://doi.org/10.5334/dsj-2023-012

Submitted: 30 November 2022
Accepted: 26 April 2023
Published: 05 June 2023

Data Science Journal is a peer-reviewed open access journal published by Ubiquity Press.