HERO-Genomics: An Ontology for Integration and Access of Multicenter Genomic Data

Mirco Cazzaro[1], Ivo G. Gut[2,3], Laura Menotti[1,*], Manuel Rueda[2,3] and Gianmaria Silvello[1]

[1]Department of Information Engineering, University of Padua, Padua, Italy
[2]Centro Nacional de Análisis Genómico (CNAG), Baldiri Reixac 4, 08028 Barcelona, Spain
[3]Universitat de Barcelona (UB), Barcelona, Spain

Abstract

The Hereditary Ontology for Genomic Data (HERO-Genomics) facilitates the structured representation of genomic information, with an initial focus on documenting genomic variations related to Amyotrophic Lateral Sclerosis (ALS) and Multiple Sclerosis (MS). The current release provides a framework for capturing specific sequence variations, focusing on Single Nucleotide Variants (SNV). HERO-Genomics is a component of the broader Hereditary Ontology (HERO), developed to model the gut-brain connection from phenoclinical and genomic viewpoints. HERO serves as the backbone of the Semantic Data Integration platform for the HEREDITARY project, utilizing the Ontology-Based Data Access (OBDA) paradigm to query heterogeneous and distributed data respecting legal constraints.

Keywords

Ontology, Genomic Data, Gut-Brain Axis, Data Modeling

1. Introduction

Genome sequencing is used to aid diagnoses, particularly of rare diseases, to inform disease treatment and progression, and to create predictive models for precision medicine [1]. The widespread availability of genomic data calls for a unified representation of genomes and trustworthy data discovery protocols. Global Alliance for Genomics and Health (GA4GH) has developed multiple initiatives to support genomic data modeling and interoperability. Among these, Phenopacket v2 [2], as a data model, and Beacon v2 [3, 4], as an API specification, are specifically designed to facilitate the sharing of phenotypic and genomic data. This communication will focus specifically on Beacon v2 and its capabilities for federated data discovery.

The Beacon v2 specification consists of two components: the *framework* and the *models*. The framework defines request and response formats, while the models specify the structure of biological data responses. Both the framework and the models are defined with JSON Schema, ensuring that API responses comply with JSON format. The Beacon v2 models consist of seven entities: *analyses*, *biosamples*, *cohorts*, *datasets*, *individuals*, *genomicVariations*, and *runs*, that serve as a template for encapsulating biomedical data. As the name suggests, all genomic data must be encapsulated within the *genomicVariations* entity. Practically, the Beacon v2 API provides the "instructions" for implementing an API, but it leaves the implementation details to the developer. To function properly, the API requires connection to a back-end where the data are stored, making the implementer responsible for the ETL (Extract, Transform, and Load) process, typically starting with a VCF file—a process also referred to as 'data beaconization.' To streamline this, the Beacon v2 Reference Implementation (B2RI) was developed

SWAT4HCLS '25: The 16th International SWAT4HCLS conference, February 24-27, 2025, Barcelona, Spain.
*Corresponding author
mirco.cazzaro@phd.unipd.it (M. Cazzaro); ivo.gut@cnag.eu (I. G. Gut); laura.menotti@unipd.it (L. Menotti); manuel.rueda@cnag.eu (M. Rueda); gianmaria.silvello@unipd.it (G. Silvello)
https://www.dei.unipd.it/~cazzaromir/ (M. Cazzaro); https://www.cnag.eu/ivo-g-gut (I. G. Gut); https://www.dei.unipd.it/~menottilau/ (L. Menotti); https://www.cnag.eu/manuel-rueda (M. Rueda); https://www.dei.unipd.it/~silvello/ (G. Silvello)
0009-0006-3856-7207 (M. Cazzaro); 0000-0001-7219-632X (I. G. Gut); 0000-0002-0676-682X (L. Menotti); 0000-0001-9280-058X (M. Rueda); 0000-0003-4970-4554 (G. Silvello)

[3]. B2RI includes a suite of tools to handle ETL for users' data. Specifically, the *beacon2-ri-tools* utility enables direct conversion of VCF data into the *genomicVariations* entity as a JSON file. This JSON file is then ingested into a MongoDB instance, allowing the included *beacon2-ri-api* to access the data without requiring on-the-fly conversion, as the data is preformatted.

Beacon v2 fully complies with the FAIR data principles, yet it is not designed to work seamlessly within semantic web ecosystems. In the HEREDITARY project, [1] our goal is to achieve interoperability that enables seamless querying across diverse but related data sources, from clinical patient records to medical imaging and genomics. We aim to unlock new research opportunities and provide a more comprehensive perspective on diseases and treatments by linking these traditionally siloed sources and enabling unified queries.

To this end, this work introduces the HEReditary Ontology for Genomic data (HERO-Genomics), an ontology allowing to query genomics data leveraging the OBDA paradigm. Indeed, OBDA requires an ontology as it defines a shared and formalized schema to interpret data, enabling consistent data representation and mapping across disparate sources. In this context, an ontology-driven approach is precious, as it ensures semantic interoperability across multicenter and heterogeneous medical data, allowing for standardized queries over distributed datasets. In this context, an ontology allows the integration of genomic data with other clinical data, e.g. clinical progression, towards federated analytics. HERO-Genomics is based on the Beacon v2 data model to employ a unified terminology that reduces ambiguity and ensures that all stakeholders interpret the data consistently.

HERO-Genomics is part of HERO, the backbone of the Semantic Data Integration platform of the HEREDITARY project, whose goal is to harmonize and link heterogeneous sources of clinical, genomics, and environmental data. HERO consists of two parts: HERO-Clinical for phenoclinical data, which is based on the BrainTeaser Ontology (BTO) [5], and HERO-Genomics. HEREDITARY exploits federated analytics and learning workflows to identify new risk factors and treatment responses focusing on brain-related diseases concerning the gut-brain axis.

The rest of this work is organized as follows: Section 2 describes the ontology design process, covering the domain requirements, the alignment with Open Biological and Biomedical Ontology Foundry (OBO) and FAIR principles, and the implementation choices. Section 3 outlines HERO-Genomics.Section 4 reports a use-case showing how HERO-Genomics can be used in practice to achieve Federated Data Access. Section 5 draws some final remarks.

2. Methodology

HERO-Genomics has been designed exploiting a co-design approach, collaborating with the medical partners and domain experts to embed their knowledge and, at the same time, to validate all the design choices. To this end, we operated iteratively, producing several (intermediate) versions of the ontology and discussing them with our domain experts. The iterative discussion process with the medical partners ensures that these newly defined concepts correctly describe the corresponding real-world concepts and guarantees the semantic quality of the ontology.

Definition of the domain requirements. In the context of the HEREDITARY project, genetic information is collected using Variant Call Format (VCF) files consisting of a variety of genomic variations such as Single Nucleotide Variants (SNVs), insertions, deletions, and structural variants, together with rich annotations [6]. Here, we will discuss the most simple of them, the Single Nucleotide Polymorphisms (SNPs). A SNP is a one-base sequence where an individual's genome varies concerning another sequence, usually called "Reference Genome". SNPs are SNVs present in a sufficiently large fraction, i.e. at least 1%, of a specific population.

VCF is a text format widely used for storing genetic variations. The format was developed for the 1000 Genome Project [2] and has been adopted in several other projects, e.g., dbSNP [3]. Each VCF

[1] https://hereditary-project.eu/
[2] https://www.internationalgenome.org/
[3] https://www.ncbi.nlm.nih.gov/snp/

file comprises a header and a body. The former provides metadata describing the file's body and keywords that optionally describe the fields used in the body. The latter is tab-separated and comprises eight mandatory columns and an unlimited number of optional columns that may record additional information about the sample.

The HERO-Genomics will answer the Beacon genomic queries provided and other queries of interest to the genomics community. Relevant Beacon queries comprise:

- Sequence Queries for the existence of a specified sequence at a given genomic position;
- Range Queries for matching variants at least overlapping with a specified region;
- GeneId Queries for returning variants affecting a gene's coding region;
- Bracket Queries for matching variants falling in a start range and end range;
- Genomic Allele Queries for matching variants with the specified allele;
- Amino Acid Change Queries for matching variants with the specified amino acid change.

2.1. Alignment with FAIR Principles and OBO Standards

HERO-Genomics adheres to the OBO principles [4] and Findable, Accessible, Interoperable, Reusable (FAIR) principles [7], [5] facilitating its adoption across heterogeneous settings. The ontology is defined in the OWL 1.2 *Common Format* and is both *open* and publicly accessible; its complete definition and detailed documentation are available at http://hereditary.dei.unipd.it/ontology/genomics/. Each entity within HERO-Genomics is identified through unique *URIs/Identifier Spaces* under the prefix https://w3id.org/hereditary/ontology/genomics/schema/. The ontology's versioning process is also described in detail in the documentation. The defined *Scope* of HERO-Genomics is *specific*: to represent genomic data from various medical centers. Following the OBO principles, we have included *Textual Definitions* for each ontology class to enhance reusability. Before introducing new relations, we reviewed existing *Relations* in the Relations Ontology (RO) to ensure none could serve the same purpose within HERO-Genomics.

Regarding the *Documented Plurality of Users* and *Commitment to Collaboration*, these aspects are central to the development and application of HERO-Genomics. HERO-Genomics was created within the HEREDITARY Project, which involves partners from multiple countries across the EU and the USA. This collaborative approach, rooted in co-design, highlights the ontology's *collaborative* nature.

The *Locus of Authority* for HERO-Genomics is vested in its developers, listed on the ontology's web page, and in the authors of this paper, which includes genomics and computer science experts. HERO-Genomics also adheres to defined *Naming Conventions*, detailed below. Lastly, the HEREDITARY consortium is committed to the ongoing *Maintenance* and updating of HERO-Genomics to ensure its relevance and accuracy.

2.2. Design Choices

We follow some basic principles when defining classes and properties to provide consistency in the HERO-Genomics.

External Referencing. Reusing entities and properties already defined in other resources enforces collaboration and data consistency [8]. External referencing is managed with annotation properties and using the Unique Resource Identifier (URI) of the term in the original thesaurus. In HERO-Genomics, external URIs are used when defining named individuals that refer to abstract concepts. On the contrary, when a new class is inserted in HERO-Genomics, it is defined within the HERO-Genomics namespace, and connected references are expressed using annotation properties.

Classes Definition and Annotation Properties. All components of the HERO-Genomics have additional information in annotation properties. All classes must have a label denoting the name and a comment, providing a brief explanation and its source (e.g., other thesauri, websites, or textbooks). The

[4] https://obofoundry.org/principles/fp-000-summary.html
[5] https://www.go-fair.org/fair-principles/

name and definition are inherited from the thesaurus if the class has an equivalent in any relevant external resource. In this case, the class comprises another annotation property called `rdfs:isDefinedby` expressing the Internationalized Resource Identifier (IRI) corresponding to the resource term of reference. If there is an additional external reference, we use the property `rdfs:seeAlso`. Most biomedical vocabularies are mapped in the Unified Medical Language System (UMLS) [6] with a unique identifier called Concept Unique Identifier (CUI) [9]. For each class that has a UMLS reference, the annotation property `dcterms:conformsTo` is instantiated with the URL of the corresponding concept.

Naming Conventions. All components must have a label and a comment. We use explanatory labels for object properties where the property range is included. In this case, the comment explains the relationship between the two classes. Concerning data properties, the label usually includes the name of the domain class so that its meaning is intuitive. A comment with the attribute description and, when available, the definition source are also included. Note that, all the HERO-Genomics components can comprise the `note` annotation property for additional remarks or business logic rules.

Usage of the Simple Knowledge Organization System (SKOS). Often we are interested in the abstract concept behind the medical term. In HERO-Genomics classism [10] is avoided for two important advantages: *i)* it dramatically reduces the number of required URIs, by not defining multiple named individuals; *ii)* it reduces the complexity of the queries. In HERO-Genomics, classification schemes that refer to abstract concepts already defined in other semantic resources, are modelled using the Simple Knowledge Organization System (SKOS) data model. [7] The SKOS data model allows storing some particular information without instantiating one individual for each patient but by simply referring to the individual already instantiated as a concept. Note that this approach prevents us from describing the peculiarities of the specific entity. However, such a design principle is employed on components that do not have this requirement, i.e. for each class referring to a set of abstract terms without any associated data or object property.

3. The HEREDITARY Ontology for Genomic Data Modelling

HERO-Genomics adheres to the Beacon v2 Data Model (v2.0), utilizing the *genomicVariations* component as the core class of the ontology. In version 2.0, this model supports the storage of genomic variations, such as SNVs, insertions, deletions, and Copy Number Variation (CNV), along with their annotations. It's worth noting that Beacon v2 is continuously evolving, with future versions expected to align with the Variant Representation Standard (VRS) [11], enabling definitions for more complex variations like Structural Variations (SV). The complete documentation of HERO-Genomics, including technical details, is available at: https://hereditary.dei.unipd.it/ontology/genomics/. The complete schema of HERO-Genomics can be found in the documentation website. [8] Figure 1 reports the schema of genomic variations and VCF files.

Genomic Variations. HERO-Genomics records systemic variations, e.g., Copy Number Change, molecular variations, and legacy variations, e.g., SNVs. Annotations such as molecular effects and amino acid changes are represented by the class "`Molecular Attribute`". Class "`Case Level Data`" stores information about phenotypic effects, clinical interpretations, and the observed zygosity, a SKOS taxonomy from the Gene Ontology [12, 13]. [9] In addition, the class links the genomic variation to the corresponding biosample ("`Biosample`"), run ("`Run`"), analysis ("`Analysis`"), and patient ("`Patient`" from HERO-Clinical. [10]) The location of the variation can be stored with the class "`Location`", which provides a representation both for sequence location, to specify the sequence interval and reference genome, and chromosome location, to specify the cytoband interval of the chromosome where the variation occurs.

[6] https://uts.nlm.nih.gov/uts/umls/home
[7] https://www.w3.org/TR/2009/REC-skos-reference-20090818/
[8] https://hereditary.dei.unipd.it/ontology/genomics/#figure1
[9] https://geneontology.org/
[10] https://hereditary.dei.unipd.it/ontology/phenoclinical/

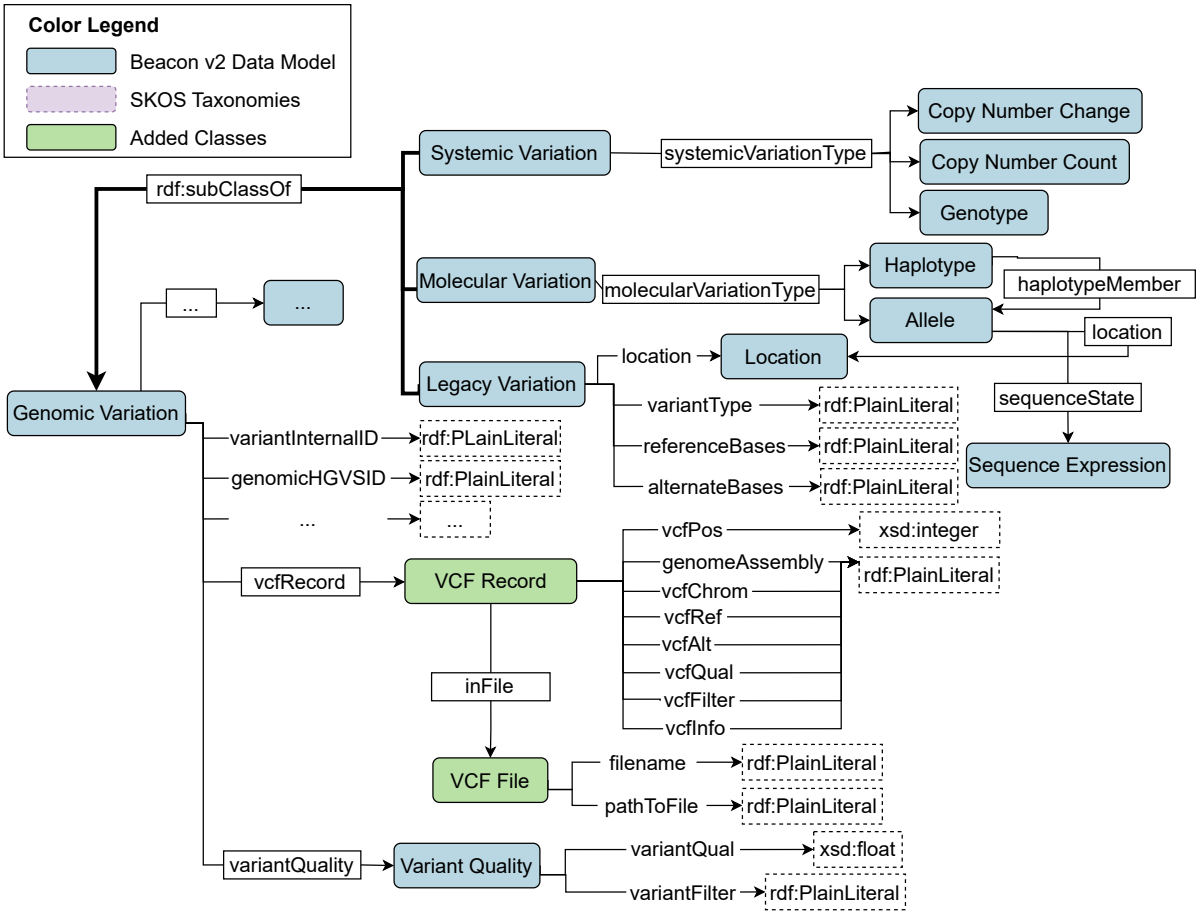

Figure 1: Genomic Variation Modeling in HERO-Genomics. HERO-Genomics records systemic variations, e.g., Copy Number Change, molecular variations, and legacy variations, e.g., SNPs. For each variation, one can store the corresponding VCF row and store additional annotations with classes "`Molecular Attribute`", "`Case Level Variant`", and "`Variant Level Data`". Classes in blue represent components from the Beacon Data Model, green was added to satisfy HEREDITARY's data requirements, and purple represents a SKOS taxonomy.

Systematic Variations comprise CNV and genotypes. CNVs represent specific DNA segments that appear in various copies among individuals. In HERO-Genomics, CNVs are represented with two classes: "`Copy Number Change`" and "`Copy Number Count`". The former describes the change in the copy number of a sequence in a genome. One can also specify the type of variation copy change exploiting a taxonomy modeled using SKOS concepts, where values are from the Experimental Factor Ontology (EFO) [14], to be specific children of "`Copy Number Assessment`". "`Copy Number Count`" represents the integer number of copies of the DNA sequence in a genome. This class stores location information with the object property `location` and the integer number of copies of the sequence with data property `variationCopies`. Systematic Variations also represent "`Genotype`", which stores the number of molecular variations with data property `moleculaVariationCount` and the molecular variations as instances of "`Genotype Member`".

Molecular Variations are variations on a contiguous molecule and can be classified into Haplotype and Allele. The former is a set of non-overlapping Allele members that co-occur on the same molecule. Thus, each haplotype is linked to its allele members with object property `haplotypeMember` ranging to class "`Allele`". The latter is a variant of the sequence of nucleotides at a particular location. Thus, class "`Allele`" has object property `location` to store the location of the allele and object property `sequenceState` to record the expression of the sequence state, represented by class "`Sequence Expression`". Sequence expressions can be "`Literal Sequence Expression`", i.e. an explicit sequence, "`Derived Sequence Expression`", i.e. a sequence that is derived from a sequence

location, “Repeated Sequence Expression”, or “Composed Sequence Expression”.

Class “Legacy Variation” represents any other genomic variation. For instance, a SNP where base "A" is mutated to "T" can be represented as a Legacy Variation, by storing its location with object property location, the variant type “SNP" with data property variantType, data property referenceBase with value “A" and alternateBase with value “T".

Variant Call Format File. One can store VCF files with classes “VCF File” and “VCF Record”. The former class comprises information about the file, such as its name and path, while the latter reports each column of the VCF row as a data property. Variant quality information, such as the quality tests performed and the variation score, can be reported using class “Variant Quality”, which comprises properties variantQual for the variation quality score and variantFilter for the performed quality tests.

4. Ontology Deployment

This section presents how the HERO-Genomics ontology can execute the Beacon queries described above. HERO-Genomics offers the same expressive capabilities as Beacon, combined with the flexibility of SPARQL Protocol and RDF Query Language (SPARQL) queries. A key advantage is the ability to query multiple genomics datasets distributed across different centers. The synthetic datasets we use here employ various data models and compression techniques. We illustrate how the OBDA paradigm [15] enables mapping the ontology to these diverse dataset sources. We employ the query in Listing 1 to show the querying process from HERO-Genomics to the local data sources and back exploiting the defined OBDA architecture.

```
PREFIX : <https://w3id.org/hereditary/ontology/schema/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT ?start ?ref ?alt
WHERE { ?variation a :GenomicVariation ;  :referenceBases ?ref ; :alternateBases ?alt ;
 :location ?seqLoc .
    ?seqLoc :sequenceInterval ?seqInt .
    ?seqInt :sequenceIntervalStart ?start .
    OPTIONAL { ?seqLoc :referenceSequence ?refName .
        FILTER (?refName = "GRCh38"^^rdf:PlainLiteral)}
    FILTER ( (?start = "9419057"^^rdf:PlainLiteral && ?ref = "C"^^rdf:PlainLiteral && ?
alt = "T"^^rdf:PlainLiteral) ||
        (?start = "719853"^^rdf:PlainLiteral && ?ref = "CAG"^^rdf:PlainLiteral && ?alt
 = "C"^^rdf:PlainLiteral) ||
        (?start = "16115625"^^rdf:PlainLiteral && ?ref = "AT"^^rdf:PlainLiteral && ?
alt = "A"^^rdf:PlainLiteral) )}
```

Listing 1: SPARQL query verifying the existence of a specified sequence at a given genomic position.

The chosen SPARQL query can be classified as a *Beacon Sequence Query*, as it verifies the existence of a specified sequence at given genomic positions. In particular, the query fetches genomic variant information and filters for reference name, starting position, reference bases, and alternate bases. The correct execution of this query should return a tuple for each position if the specified sequence exists.

The datasets used in this case study are publicly available genomic datasets containing variant information and zygosity sampling. We utilize curated synthetic datasets from the Common Infrastructure for National Cohorts in Europe, Canada, and Africa (CINECA) project website: CINECA synthetic cohort NA Canada CHILD v1; [11] [12] CINECA synthetic cohort Europe CH SIB; [13] CINECA synthetic cohort Africa H3ABioNet v1. [14]

[11] https://zenodo.org/records/5122832
[12] https://www.cineca-project.eu/
[13] https://zenodo.org/records/4955933
[14] https://zenodo.org/records/5082689

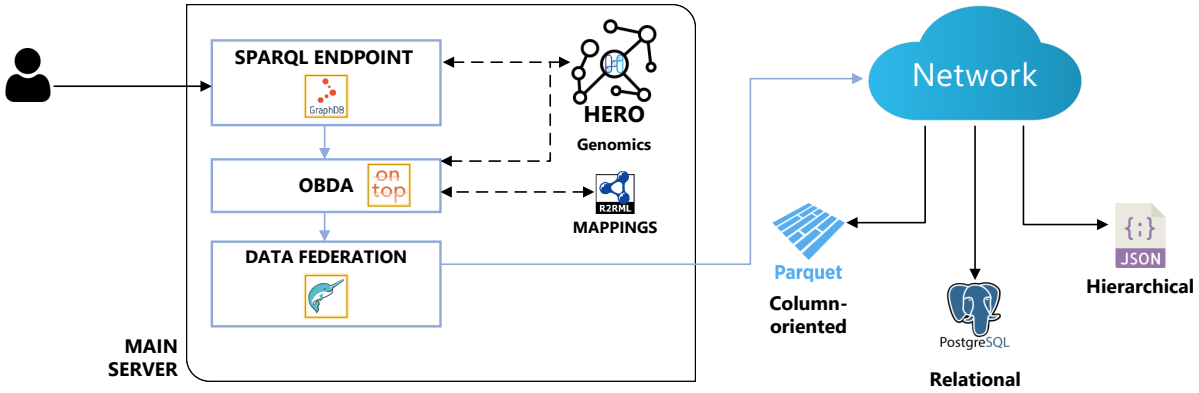

Figure 2: Federated software architecture implementing the OBDA paradigm. The framework comprises a SPARQL endpoint, an OBDA component exploiting the ontology and mapping definitions, and a Data Federation system that collects and virtualizes data from three diverse data sources: a column-oriented file, a hierarchical DB and a relational DB.

We identified three key entities within the unstructured VCF schema: "VARIANT," "VARIANT INFO," and "SAMPLE." The VCF datasets were then serialized across three different Database Management Systems (DBMSs): a PostgreSQL DB, a JavaScript Object Notation (JSON) hierarchical DB, and an Apache Parquet column-oriented DB. These distinct data models typically cannot be queried seamlessly together, making this an ideal use case to demonstrate the capabilities of the OBDA paradigm in enabling unified access across heterogeneous data sources.

The primary goal of the Federated OBDA architecture is to provide access to data through an ontology, mimicking the behavior of an Resource Description Framework (RDF)-based knowledge bases and replicating graph databases' key features. Figure 2 provides an overview of different software components are plugged into our use case scenario. The three data sources are connected to the data federation system, which primarily collects and combines data under a comprehensive virtual relational schema. We mapped this schema with the ontology vocabulary using the Ontop native mapping language. [15] An example of mapping definition is presented in Listing 2.

```
mappingId MAPID-868c2fe1dffc4fa2b01888540fc4639b
target   :chr{chrom}{pos}{ref}{alt} :variantAlternativeID :chr{chrom}{pos}{ref}{alt}
    AlternativeID ; :referenceBases {ref}^^rdf:PlainLiteral ; :alternateBases {alt}^^rdf:
    PlainLiteral . :chr{chrom}{pos}{ref}{alt}AlternativeID :valueID {id}^^rdf:PlainLiteral
     .
source    SELECT chrom, pos, "ref", alt, id FROM public_json_parquet_psql.variants;
```

Listing 2: A mapping definition in Ontop native mapping language in this case mapping a SPARQL query to a SQL query. The mapping definition comprises three fields: the mapping id, the target triples in the ontology that must be mapped, and the source SQL query that retrieves the data from a relational database.

To implement the principles of the OBDA paradigm, target triples in the ontology are parameterized with placeholders enclosed in curly brackets. These placeholders correspond to fields in the relational schema referenced by the associated SQL source query. In our use case, the diverse and heterogeneous nature of the data sources poses challenges beyond the standard capabilities of the OBDA paradigm, which traditionally supports only a single relational data source. However, multiple data sources can be federated under a virtual schema, enabling unified and comprehensive access to data. Our approach integrates three data sources into a virtual schema that supports SQL queries. This abstraction layer, often called a virtual data lake, bridges the ontology and the local data sources using OBDA mappings. Each row retrieved from the virtual data lake produces a set of virtual triples. These triples are generated

[15] https://ontop-vkg.org/guide/concepts.html#mappings

start	ref	alt
16115625	AT	A
9419057	C	T
719853	CAG	C

Table 1
Retrieved result set from Ontop after execution of query represented in Listing 1.

by matching queried fields with the corresponding placeholders defined in the mapping. For instance, as illustrated in Listing 2, target triples expressed in Turtle notation can capture variant information such as the genomic variation ID, the reference base, and the alternate base. The parameters enclosed in brackets are dynamically extracted from the source query. The OBDA component utilizes these mapping definitions and the ontology to rewrite the original SPARQL query. This process allows for including non-explicit results and translates the query into SQL, ensuring adequate access to the underlying data.

In our setup, we employed Dremio [16] as virtual data lake federation system, while GraphDB [17] provided both the SPARQL endpoint and the OBDA tool (i.e., Ontop [16]). GraphDB and Dremio were instantiated in a local environment. At the same time, the PostgreSQL data source and Network Attached Storage (NAS) repositories, i.e., Apache Parquet DB and the JSON-based DB, were remote. The data sources from these sources were virtualized in Dremio by delineating virtual views, and semantic mappings have been defined between the virtual schema and HERO-Genomics. The results of Query 1 are in Table 1.

In blue (●) we highlighted the row coming from the PostgreSQL; in yellow (●) the row coming from the Apache Parquet dataset; in green (●) the row coming from the JSON dataset. This demonstrates how the data federation system enables querying of three heterogeneous and remote data sources. We can interpret relationships between the data and construct meaningful queries by employing HERO-Genomics as a unified semantic layer. The resulting dataset is a unified collection of triples, seamlessly integrated despite originating from sources that could not otherwise be queried together.

5. Conclusion

This work presents the HERO-Genomics, an ontology enabling the communication between Beacon v2 and RDF. HERO-Genomics follows the Beacon Data Model, a *de facto* standard adopted by the genomics community to enable federated discovery of genomic data. Our implementation of HERO-Genomics within the HEREDITARY project's Semantic Data Integration platform shows its potential for harmonizing diverse data types, including clinical, genomic, and environmental information. This integration paves the way for more comprehensive research in brain-related diseases and disorders of the gut-brain axis. Moreover, aligning with the FAIR principles, HERO-Genomics promotes data discoverability, accessibility, and usability, which are crucial for advancing precision medicine. Through the example use-case in federated data access, we demonstrated how HERO-Genomics can facilitate federated querying of genomics data, even across multiple, heterogeneous datasets. This capability is pivotal for large-scale research collaborations and cross-institutional studies, where seamless access to comprehensive, interoperable data can accelerate scientific discovery.

Future phases will consist of further refinements to HERO-Genomics for expanded genomics data capabilities. HERO-Genomics is part of the HERO, which is the backbone of the Semantic Data Integration platform of the HEREDITARY project, exploiting the OBDA technology to query, aggregate, and join large heterogeneous data in a distributed manner using a unique query language, i.e., SPARQL. HERO will be expanded to include aspects of the gut-brain interplay, focusing on developing a gut microbiome ontology, an area currently with limited foundational work or related models. This approach will ensure that HERO remains robust and scalable, allowing it to capture complex biomedical data across various brain-related conditions and emerging research areas.

[16] https://www.dremio.com/
[17] https://graphdb.ontotext.com/

Acknowledgments

This project has received funding from the HEREDITARY Project as part of the European Union's Horizon Europe research and innovation programme under grant agreement No GA 101137074.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT, and Grammarly in order to: Improve writing style, and Formatting assistance. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] K. J. Karczewski, M. P. Snyder, Integrative omics for health and disease, Nature Reviews Genetics 19 (2018) 299–310. URL: https://doi.org/10.1038/nrg.2018.4.

[2] M. S. Ladewig, J. O. B. Jacobsen, A. H. Wagner, D. Danis, B. El Kassaby, M. Gargano, T. Groza, M. Baudis, R. Steinhaus, D. Seelow, N. E. Bechrakis, C. J. Mungall, P. N. Schofield, O. Elemento, L. Smith, J. A. McMurry, M. Munoz-Torres, M. A. Haendel, P. N. Robinson, GA4GH Phenopackets: A Practical Introduction, Advanced Genetics 4 (2023) 2200016. URL: https://doi.org/10.1002/ggn2.202200016.

[3] M. Rueda, R. Ariosa, M. Moldes, J. Rambla, Beacon v2 Reference Implementation: a toolkit to enable federated sharing of genomic and phenotypic data, Bioinformatics 38 (2022) 4656–4657. URL: https://doi.org/10.1093/bioinformatics/btac568.

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

[5] G. Faggioli, L. Menotti, S. Marchesin, A. Chió, A. Dagliati, M. de Carvalho, M. Gromicho, U. Manera, E. Tavazzi, G. M. Di Nunzio, G. Silvello, N. Ferro, An extensible and unifying approach to retrospective clinical data modeling: the brainteaser ontology, Journal of Biomedical Semantics 15 (2024) 16. URL: https://doi.org/10.1186/s13326-024-00317-y.

[6] P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin, . G. P. A. Group, The variant call format and VCFtools, Bioinformatics 27 (2011) 2156–2158. URL: https://doi.org/10.1093/bioinformatics/btr330.

[7] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, The fair guiding principles for scientific data management and stewardship, Scientific Data 3 (2016) 160018. URL: https://doi.org/10.1038/sdata.2016.18.

[8] E. Simperl, Reusing ontologies on the semantic web: A feasibility study, Data & Knowledge Engineering 68 (2009) 905–925. URL: https://doi.org/10.1016/j.datak.2009.02.002.

[9] O. Bodenreider, The Unified Medical Language System (UMLS): integrating biomedical terminology, Nucleic Acids Research 32 (2004) 267–270. URL: https://doi.org/10.1093/nar/gkh061.

[10] D. Allemang, J. Hendler, F. Gandon, Good and bad modeling practices, in: Semantic Web for the Working Ontologist: Effective Modeling for Linked Data, RDFS, and OWL, Association for Computing Machinery, New York, NY, USA, 2020, pp. 436–440. URL: https://doi.org/10.1145/3382097.3382113.

[11] A. H. Wagner, L. Babb, G. Alterovitz, M. Baudis, M. Brush, D. L. Cameron, M. Cline, M. Griffith, O. L. Griffith, S. E. Hunt, D. Kreda, J. M. Lee, S. Li, J. Lopez, E. Moyer, T. Nelson, R. Y. Patel, K. Riehle, P. N. Robinson, S. Rynearson, H. Schuilenburg, K. Tsukanov, B. Walsh, M. Konopko, H. L. Rehm, A. D. Yates, R. R. Freimuth, R. K. Hart, The ga4gh variation representation specification: A computational framework for variation representation and federated identification, Cell Genomics 1 (2021) 100027. doi:10.1016/j.xgen.2021.100027.

[12] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, G. Sherlock, Gene ontology: tool for the unification of biology, Nature Genetics 25 (2000) 25–29. URL: https://doi.org/10.1038/75556.

[13] T. G. O. Consortium, S. A. Aleksander, J. Balhoff, S. Carbon, J. M. Cherry, H. J. Drabkin, D. Ebert, M. Feuermann, P. Gaudet, N. L. Harris, D. P. Hill, R. Lee, H. Mi, S. Moxon, C. J. Mungall, A. Muruganugan, T. Mushayahama, P. W. Sternberg, P. D. Thomas, K. Van Auken, J. Ramsey, D. A. Siegele, R. L. Chisholm, P. Fey, M. C. Aspromonte, M. V. Nugnes, F. Quaglia, S. Tosatto, M. Giglio, S. Nadendla, G. Antonazzo, H. Attrill, G. dos Santos, S. Marygold, V. Strelets, C. J. Tabone, J. Thurmond, P. Zhou, S. H. Ahmed, P. Asanitthong, D. Luna Buitrago, M. N. Erdol, M. C. Gage, M. Ali Kadhum, K. Y. C. Li, M. Long, A. Michalak, A. Pesala, A. Pritazahra, S. C. C. Saverimuttu, R. Su, K. E. Thurlow, R. C. Lovering, C. Logie, S. Oliferenko, J. Blake, K. Christie, L. Corbani, M. E. Dolan, H. J. Drabkin, D. P. Hill, L. Ni, D. Sitnikov, C. Smith, A. Cuzick, J. Seager, L. Cooper, J. Elser, P. Jaiswal, P. Gupta, P. Jaiswal, S. Naithani, M. Lera-Ramirez, K. Rutherford, V. Wood, J. L. De Pons, M. R. Dwinell, G. T. Hayman, M. L. Kaldunski, A. E. Kwitek, S. J. F. Laulederkind, M. A. Tutaj, M. Vedi, S.-J. Wang, P. D'Eustachio, L. Aimo, K. Axelsen, A. Bridge, N. Hyka-Nouspikel, A. Morgat, S. A. Aleksander, J. M. Cherry, S. R. Engel, K. Karra, S. R. Miyasato, R. S. Nash, M. S. Skrzypek, S. Weng, E. D. Wong, E. Bakker, T. Z. Berardini, L. Reiser, A. Auchincloss, K. Axelsen, G. Argoud-Puy, M.-C. Blatter, E. Boutet, L. Breuza, A. Bridge, C. Casals-Casas, E. Coudert, A. Estreicher, M. Livia Famiglietti, M. Feuermann, A. Gos, N. Gruaz-Gumowski, C. Hulo, N. Hyka-Nouspikel, F. Jungo, P. Le Mercier, D. Lieberherr, P. Masson, A. Morgat, I. Pedruzzi, L. Pourcel, S. Poux, C. Rivoire, S. Sundaram, A. Bateman, E. Bowler-Barnett, H. Bye-A-Jee, P. Denny, A. Ignatchenko, R. Ishtiaq, A. Lock, Y. Lussi, M. Magrane, M. J. Martin, S. Orchard, P. Raposo, E. Speretta, N. Tyagi, K. Warner, R. Zaru, A. D. Diehl, R. Lee, J. Chan, S. Diamantakis, D. Raciti, M. Zarowiecki, M. Fisher, C. James-Zorn, V. Ponferrada, A. Zorn, S. Ramachandran, L. Ruzicka, M. Westerfield, The Gene Ontology knowledgebase in 2023, Genetics 224 (2023) iyad031. URL: https://doi.org/10.1093/genetics/iyad031.

[14] J. Malone, E. Holloway, T. Adamusiak, M. Kapushesky, J. Zheng, N. Kolesnikov, A. Zhukova, A. Brazma, H. Parkinson, Modeling sample variables with an Experimental Factor Ontology, Bioinformatics 26 (2010) 1112–1118. URL: https://doi.org/10.1093/bioinformatics/btq099.

[15] D. Calvanese, G. D. Giacomo, D. Lembo, M. Lenzerini, R. Rosati, Ontology-Based Data Access and Integration, Springer New York, New York, NY, 2018, pp. 2590–2596. URL: https://doi.org/10.1007/978-1-4614-8265-9_80667.

[16] D. Calvanese, B. Cogrel, E. G. Kalayci, S. Komla-Ebri, R. Kontchakov, D. Lanti, M. Rezk, M. Rodriguez-Muro, G. Xiao, OBDA with the ontop framework, in: Proc. of the 23rd Italian Symposium on Advanced Database Systems (SEBD 2015), Gaeta, Italy, June 14-17, 2015, Curran Associates, Inc., 2015, pp. 296–303.