University of Murcia
Faculty of Biology
Master's Degree in Bioinformatics

Semantic Representation and Analysis of Genes Associated with Muscular Dystonia

Carlos C. Ureña Mateo
Academic Year 2024/2025
University of Murcia

Contents

List of Figures

List of Tables

1 Introduction and Objectives

Dystonia comprises a heterogeneous spectrum of movement disorders characterized by involuntary muscle contractions that produce abnormal postures, twisting movements, or tremor-like manifestations. These motor symptoms may affect a single region or extend across multiple anatomical areas, and they often display a characteristic, patterned distribution that evolves over time. Clinically, dystonia is classified according to its topography (focal, segmental, generalized), age at onset, and temporal behavior, with isolated dystonia being defined as the sole or predominant neurological feature, and combined dystonia occurring alongside other movement abnormalities such as myoclonus or parkinsonism [2, 1].

This condition has a diverse etiological basis, ranging from acquired causes—such as perinatal brain injury or drug-induced damage—to genetic variants affecting critical neuronal pathways. In particular, isolated forms of dystonia are frequently associated with pathogenic mutations, even in the absence of a clear family history. Genetic contributions are especially relevant in early-onset cases, with multiple loci now recognized as causative or contributory. Notable among these are TOR1A, PANK2, ATP1A3, PRRT2, GCH1, THAP1, SGCE and ANO3, which encode proteins involved in neurotransmission, ion homeostasis, transcriptional regulation, and mitochondrial dynamics [4, 1].

The landscape of dystonia genetics has expanded considerably with the application of next-generation sequencing (NGS) technologies, allowing the discovery of novel variants across idiopathic and familial cases. While whole-genome and whole-exome sequencing have revealed broad mutational profiles, targeted sequencing approaches remain especially valuable in the diagnostic evaluation of clinically heterogeneous dystonia patients, offering greater coverage of known loci and facilitating variant interpretation [4].

Functionally, the genes implicated in dystonia are diverse and intersect with multiple biological systems. These include pathways of dopaminergic transmission, calcium channel regulation, heavy metal detoxification, and energy metabolism. Disruptions in these pathways may lead to shared neurobiological dysfunctions that underlie different dystonia phenotypes. Understanding such convergences has become a key objective in the search for mechanism-based therapeutic targets [3].

This project proposes a semantic modeling approach to represent structured knowledge about genes implicated in muscular dystonia. Through the use of Resource Description Framework (RDF) and ontology-based annotation, the aim is to create a machine-readable, interoperable dataset that integrates genomic and clinical data. By enabling precise queries and inference over this semantic model, the project seeks to support advanced biomedical analyses and foster data reuse in the context of rare neurogenetic diseases.

2 Materials and Methods

2.1 Data Sources and Gene Selection

The construction of the RDF dataset began with a targeted review of the biomedical literature. Specifically, a peer-reviewed publication focusing on the molecular underpinnings of muscular dystonia was used to identify a set of eight genes recurrently associated with

this condition [1]. Each gene was selected based on evidence of its involvement in the clinical phenotype of dystonia, including causative mutations and associated syndromes.

Gene	**Genetic Alteration**	**Associated Syndrome**
TOR1A	c.907_909delGAG (p.Glu303del)	DYT1 early-onset generalized dystonia
PANK2	c.1133A>G (p.Asp378Gly)	Pantothenate kinase-associated neurodegeneration
ATP1A3	c.1871C>T (p.Thr624Met)	Rapid-onset dystonia-parkinsonism
PRRT2	c.859G>A (p.Ala287Thr)	Paroxysmal kinesigenic dyskinesia
GCH1	c.317C>T (p.Thr106Ile)	Dopa-responsive dystonia
THAP1	c.97T>G (p.Cys33Gly)	Early-onset dystonia type 6
SGCE	c.109+5G>C (splicing)	Myoclonus-dystonia
ANO3	c.1409G>A (p.Gly470Glu)	Craniocervical dystonia

Table 1: Genetic alterations and associated syndromes extracted from the literature.

To enrich and validate the dataset, additional annotations were retrieved from established biomedical resources. The NCBI Gene database was consulted to obtain the standardized gene symbols and chromosomal locations. The UniProt database provided information on the protein products encoded by these genes, including their names and accession numbers. Clinical associations were curated from the OMIM database, and disease identifiers were aligned with the MONDO Disease Ontology to enable semantic interoperability.

Gene	**Protein**	**Chromosome**	**OMIM**	**MONDO**	**UniProt**
TOR1A	Torsin-1A	9q34.11	128100	MONDO:0007492	O14656
PANK2	Pantothenate kinase 2	20p13	234200	MONDO:0009319	Q9BZ23
ATP1A3	Na^+/K^+-ATPase alpha-3	19q13.2	128235	MONDO:0007496	P13637
PRRT2	Proline-rich TM protein 2	16p11.2	128200	MONDO:0044202	Q7Z6L0
GCH1	GTP cyclohydrolase 1	14q22.2	128230	MONDO:0016812	P30793
THAP1	THAP domain protein 1	8p11.21	605909	MONDO:0100016	Q9NVV9
SGCE	Epsilon-sarcoglycan	7q21.3	159900	MONDO:0000903	O43556
ANO3	Anoctamin-3	11p14.3-p14.2	615034	MONDO:0011886	Q9BYT9

Table 2: Additional gene-level annotations obtained from external databases (NCBI, UniProt, OMIM, MONDO).

2.2 Ontology Design and RDF Dataset Construction

To formally represent the biological knowledge associated with muscular dystonia, a custom ontology was developed using the Resource Description Framework (RDF) expressed in Turtle syntax. This ontology captures the relationships between genes and clinical syndromes, allowing semantic integration with external biomedical knowledge bases and facilitating advanced querying via SPARQL.

The construction of the RDF dataset was performed in successive stages: beginning with the declaration of prefixes, followed by the definition of ontology terms (classes and properties), and concluding with the instantiation of real-world entities as individuals within the graph.

2.2.1 Prefix Declaration

The RDF dataset begins with a structured declaration of **prefixes**, which serve as semantic abbreviations for full URIs. This strategy enhances both the readability and maintainability of the Turtle code, while promoting alignment with external ontologies and data sources. The prefixes used are listed below:

Prefix Declarations

```
@prefix dist_r: <https://um.es/data/dystonia/> .       # Instances
@prefix dist_o: <https://dystonia_ontology.um.es/> .   # Ontology terms
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix ncbi: <https://www.ncbi.nlm.nih.gov/gene/> .
@prefix uniprot: <https://www.uniprot.org/uniprot/> .
@prefix omim: <https://omim.org/entry/> .
@prefix mondo: <http://purl.obolibrary.org/obo/MONDO_> .
@prefix doid: <http://purl.obolibrary.org/obo/DOID_> .
@prefix ro: <http://purl.obolibrary.org/obo/RO_> .
@prefix sio: <http://semanticscience.org/resource/> .
```

A modular and unified URI strategy was adopted to ensure semantic clarity and facilitate dereferenceability. Two distinct base namespaces were defined:

- `dist_o:` for the definition of ontology terms such as classes and properties.
- `dist_r:` for the instantiation of domain-specific entities including genes and syndromes.

This separation aligns with best practices in ontology engineering, distinguishing the **terminological component** (TBox), which defines the vocabulary, from the **assertional component** (ABox), which contains factual assertions about real-world entities. It also guarantees compatibility with Linked Data publication platforms like Trifid, which require a clear and dereferenceable URI base.

The standard RDF and OWL prefixes (`rdf:`, `rdfs:`, `owl:`, `xsd:`) are essential for defining core elements of the RDF graph, such as types, labels, comments, and data types. Prefixes for biomedical databases (`ncbi:`, `uniprot:`, `omim:`, `mondo:`) enable cross-referencing with authoritative resources, thereby enriching the dataset with curated biological knowledge.

Finally, ontology alignment prefixes (`doid:`, `ro:`, `sio:`) are used to link local terms with semantically equivalent properties and classes from widely accepted biomedical ontologies. This facilitates data interoperability and promotes reuse within the broader Semantic Web ecosystem.

2.2.2 Ontology Terms: Classes and Properties

After defining the prefixes, the ontology terms were declared to formally structure the semantic domain of muscular dystonia. These terms include the core **classes** and **properties** needed to represent genes and syndromes, as well as the relevant biological relationships between them.

Classes. Two main classes were defined to represent the key biomedical entities in the dataset: genes and clinical syndromes. Each class includes a human-readable label and a semantic alignment to an external ontology using `owl:equivalentClass`.

Class Declarations

```
dist_o:Gene rdf:type owl:Class ;
    rdfs:label "Gene" ;
    owl:equivalentClass <http://purl.obolibrary.org/obo/SO_0000704> .

dist_o:Syndrome rdf:type owl:Class ;
    rdfs:label "Syndrome" ;
    owl:equivalentClass <http://purl.obolibrary.org/obo/DOID_225> .
```

These equivalences align the local ontology with standard biomedical vocabularies, namely the Sequence Ontology (SO) for genes and the Disease Ontology (DOID) for syndromes. This alignment enhances interoperability and enables integration with other semantic datasets in the biomedical domain.

Properties. The ontology defines three custom properties within the `dist_o:` namespace and reuses one external property from the Relation Ontology (RO). All properties include human-readable labels and explanatory comments, and some are semantically aligned with external ontologies via `owl:equivalentProperty`.

Property Declarations

```
dist_o:hasAlteration rdf:type owl:DatatypeProperty ;
    rdfs:label "has alteration" ;
    rdfs:comment "Describes a genetic mutation or variation in the gene." ;
    owl:equivalentProperty <http://semanticscience.org/resource/SIO_001114> .

dist_o:isAssociatedWith rdf:type owl:ObjectProperty ;
    rdfs:label "is associated with" ;
    rdfs:comment "Relates a gene to a clinical syndrome or disease." ;
    owl:equivalentProperty <http://semanticscience.org/resource/SIO_000628> .

dist_o:locatedOnChromosome rdf:type owl:DatatypeProperty ;
    rdfs:label "located on chromosome" ;
    rdfs:comment "Specifies the chromosomal location of a gene." ;
    rdfs:domain dist_o:Gene ;
    rdfs:range xsd:string .

ro:0002205 rdf:type owl:ObjectProperty ;
    rdfs:label "encodes" ;
    rdfs:comment "Relates a gene to the protein it encodes." .
```

This property model enables a precise semantic representation of genetic information relevant to dystonia, while remaining compatible with external biomedical ontologies such as SIO and RO. The inclusion of domain and range constraints for datatype properties further improves the expressiveness and validation capabilities of the ontology.

2.2.3 Individuals and Data Instantiation

Once the ontology schema was established, individual instances were created to represent real-world entities relevant to muscular dystonia. These individuals were declared under the `dist_r:` namespace and include genes and their associated syndromes. Each gene instance was semantically enriched through references to external biomedical resources such as NCBI Gene, UniProt, and OMIM following the principles of Linked Data. Syndromes were additionally linked to MONDO terms.

For example, the gene `TOR1A`, known to be involved in early-onset generalized dystonia, was instantiated as follows:

Individual Example: TOR1A and DYT1

```
dist_r:TOR1A rdf:type dist_o:Gene ;
    rdfs:label "TOR1A" ;
    owl:sameAs ncbi:1861 ;
    ro:0002205 uniprot:O14656 ;
    dist_o:locatedOnChromosome "9q34.11"^^xsd:string ;
    dist_o:hasAlteration "c.907_909delGAG (p.Glu303del)" ;
    rdfs:seeAlso omim:128100 ;
    dist_o:isAssociatedWith dist_r:DYT1 .

uniprot:O14656 rdfs:label "Torsin-1A" .

dist_r:DYT1 rdf:type dist_o:Syndrome ;
    rdfs:label "DYT1 early-onset generalized dystonia" ;
    owl:sameAs mondo:0007492 .
```

This RDF encoding includes the following semantic elements:

- `rdf:type` declares the instance's class (e.g., `dist_o:Gene` or `dist_o:Syndrome`).
- `rdfs:label` provides a human-readable name for the entity.
- `owl:sameAs` links the entity to an authoritative external resource.
- `ro:0002205` connects the gene to the UniProt protein it encodes.
- `dist_o:locatedOnChromosome` and `dist_o:hasAlteration` provide molecular-level information.
- `dist_o:isAssociatedWith` establishes the clinical link to a syndrome.

The syndrome instance `dist_r:DYT1` is defined immediately after the gene, since it acts as the object of the `dist_o:isAssociatedWith` relation. While syndromes could be grouped separately, placing them adjacent to their related genes improves readability and facilitates contextual navigation within the RDF graph.

This instantiation pattern was systematically applied to all eight genes in the dataset, including `PANK2`, `ATP1A3`, `PRRT2`, `GCH1`, `THAP1`, `SGCE`, and `ANO3`, each of which is semantically linked to its associated syndrome through well-defined RDF triples. All instances were enriched with curated annotations and dereferenceable URIs, supporting reasoning, federated SPARQL querying, and integration into the broader Semantic Web ecosystem.

2.3 Validation and Format Conversion of the RDF Dataset

After the construction of the RDF dataset in Turtle syntax, a format transformation and validation phase was performed to ensure both semantic integrity and technical interoperability with Semantic Web tools and standards.

Format conversion. The dataset was first converted from Turtle to RDF/XML using the EasyRDF Converter (https://www.easyrdf.org/converter), a tool that supports bidirectional transformation across all major RDF serializations, including RDF/XML, N-Triples, and N-Quads. This conversion was necessary to meet the input requirements of several triple stores, such as Blazegraph, which often expect RDF/XML.

Syntax validation. The converted RDF/XML file was then validated using the official W3C RDF Validator (https://www.w3.org/RDF/Validator/). This tool verifies structural correctness, adherence to RDF syntax rules, and semantic consistency of the graph. Figure 1 shows a screenshot confirming the successful validation of the dataset.

Your RDF document validated successfully.

Triples of the Data Model

Number	Subject	Predicate	Object
1	https://dystonia_ontology.um.es/Gene	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2002/07/owl#Class
2	https://dystonia_ontology.um.es/Gene	http://www.w3.org/2000/01/rdf-schema#label	"Gene"
3	https://dystonia_ontology.um.es/Gene	http://www.w3.org/2002/07/owl#equivalentClass	http://purl.obolibrary.org/obo/SO_0000704
4	https://dystonia_ontology.um.es/Syndrome	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2002/07/owl#Class
5	https://dystonia_ontology.um.es/Syndrome	http://www.w3.org/2000/01/rdf-schema#label	"Syndrome"
6	https://dystonia_ontology.um.es/Syndrome	http://www.w3.org/2002/07/owl#equivalentClass	http://purl.obolibrary.org/obo/DOID_225
7	https://dystonia_ontology.um.es/hasAlteration	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2002/07/owl#DatatypeProperty
8	https://dystonia_ontology.um.es/hasAlteration	http://www.w3.org/2000/01/rdf-schema#label	"has alteration"
9	https://dystonia_ontology.um.es/hasAlteration	http://www.w3.org/2000/01/rdf-schema#comment	"Describes a genetic mutation or variation in the gene."
10	https://dystonia_ontology.um.es/hasAlteration	http://www.w3.org/2002/07/owl#equivalentProperty	http://semanticscience.org/resource/SIO_001114
11	https://dystonia_ontology.um.es/isAssociatedWith	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2002/07/owl#ObjectProperty
12	https://dystonia_ontology.um.es/isAssociatedWith	http://www.w3.org/2000/01/rdf-schema#label	"is associated with"
13	https://dystonia_ontology.um.es/isAssociatedWith	http://www.w3.org/2000/01/rdf-schema#comment	"Relates a gene to a clinical syndrome or disease."
14	https://dystonia_ontology.um.es/isAssociatedWith	http://www.w3.org/2002/07/owl#equivalentProperty	http://semanticscience.org/resource/SIO_000628
15	https://dystonia_ontology.um.es/locatedOnChromosome	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2002/07/owl#DatatypeProperty
16	https://dystonia_ontology.um.es/locatedOnChromosome	http://www.w3.org/2000/01/rdf-schema#label	"located on chromosome"
17	https://dystonia_ontology.um.es/locatedOnChromosome	http://www.w3.org/2000/01/rdf-schema#comment	"Specifies the chromosomal location of a gene."

Figure 1: Confirmation of the successful validation of the RDF dataset.

The validation confirmed that all prefix declarations were properly defined, URIs were correctly formed, and that all triples followed the subject–predicate–object structure.

Figure 2 shows the graphical representation of the RDF graph generated by the validator. While the visualization may be dense due to the number of nodes and edges, it serves to confirm that the semantic graph has been constructed successfully and that all interlinked entities are correctly instantiated.

Together, these steps ensured that the RDF dataset is not only syntactically correct and semantically coherent, but also technically interoperable and ready for deployment. The dataset is now suitable for integration into triple stores, enabling SPARQL querying, Linked Data publication, and adherence to FAIR data principles.

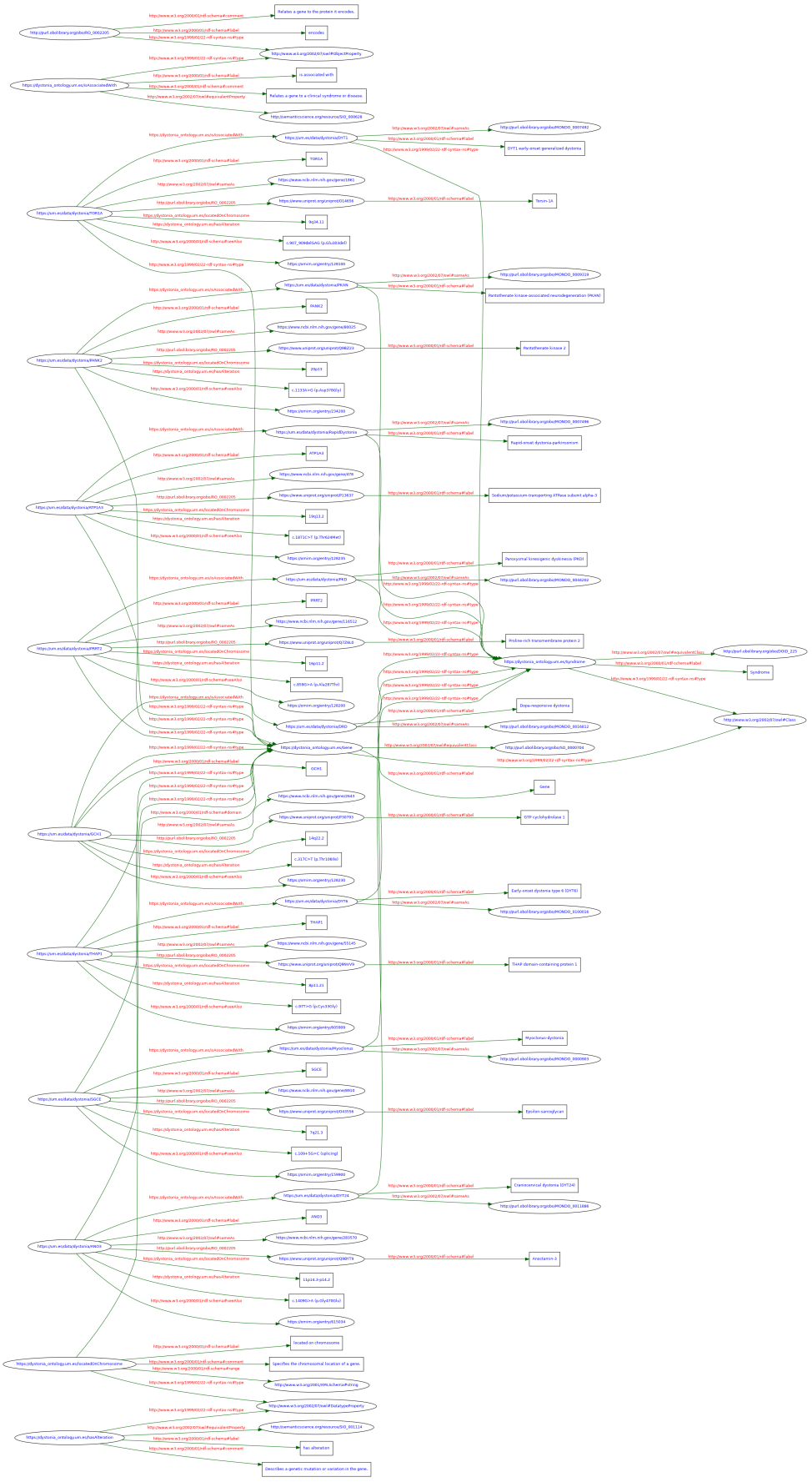

Figure 2: Graphical rendering of the RDF dataset generated by the W3C RDF Validator.

2.4 SPARQL

To interrogate the RDF knowledge graph constructed for this project, a set of five SPARQL queries was designed to address biologically meaningful questions, such as gene–syndrome associations, mutation types, and cross-references to external protein identifiers. These queries were first developed and tested using the Blazegraph Workbench graphical interface, available at http://dayhoff.inf.um.es:3050, where the dataset was loaded into a dedicated namespace named `dystonia`.

To complement this exploration and illustrate how semantic queries can be integrated into automated workflows, the same queries were also implemented in R using the `SPARQL` package (see Appendix B for the full code). The script connects to the same Blazegraph endpoint at http://dayhoff.inf.um.es:3050/blazegraph/namespace/dystonia/sparql and was validated to replicate the results obtained via the Workbench. Although not strictly required, this R implementation serves as a reproducible example of how FAIR-compliant datasets can be queried programmatically.

3 SPARQL queries

Query 1: Genes with splicing-related alterations

This query aims to identify genes associated with dystonia whose documented mutations affect RNA splicing. Splicing-related mutations are particularly relevant in clinical genomics, as they can interfere with post-transcriptional regulation and frequently underlie complex neurological phenotypes. To retrieve such genes, the query filters for the presence of the keyword "splicing" within the textual description of each genetic alteration.

SPARQL Query: Genes with splicing-related alterations

```
PREFIX dist_r: <https://um.es/data/dystonia/>
PREFIX dist_o: <https://dystonia_ontology.um.es/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT DISTINCT ?geneLabel ?chromosome ?alteration ?syndromeLabel
WHERE {
  ?gene a dist_o:Gene ;
        rdfs:label ?geneLabel ;
        dist_o:hasAlteration ?alteration ;
        dist_o:locatedOnChromosome ?chromosome ;
        dist_o:isAssociatedWith ?syndrome .

  ?syndrome rdfs:label ?syndromeLabel .

  FILTER CONTAINS(LCASE(STR(?alteration)), "splicing")
}
ORDER BY ?geneLabel
```

Query results:

Gene	Chromosome	Alteration	Syndrome
SGCE	7q21.3	c.109+5G>C (splicing)	Myoclonus-dystonia

Table 3: Gene affected by a splicing-related mutation and its associated clinical syndrome.

Query 2: Clinical syndromes associated with the gene *ANO3*

This query focuses on the gene *ANO3*, a known contributor to certain forms of dystonia. The query retrieves its associated clinical syndrome, the MONDO disease ontology identifier for standardized referencing, and complementary information including the gene name and chromosomal location. This allows for contextualizing the gene within both genomic and phenotypic frameworks.

SPARQL Query: Syndromes associated with ANO3

```
PREFIX dist_r: <https://um.es/data/dystonia/>
PREFIX dist_o: <https://dystonia_ontology.um.es/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>

SELECT DISTINCT ?geneLabel ?chromosome ?syndromeLabel ?mondoID
WHERE {
  dist_r:ANO3 rdfs:label ?geneLabel ;
              dist_o:locatedOnChromosome ?chromosome ;
              dist_o:isAssociatedWith ?syndrome .

  ?syndrome rdfs:label ?syndromeLabel ;
            owl:sameAs ?mondoID .
}
```

Query results:

Gene	Chromosome	Syndrome	MONDO ID
ANO3	11p14.3-p14.2	Craniocervical dystonia (DYT24)	MONDO_0011886

Table 4: Syndrome associated with the gene *ANO3*, including chromosomal location and MONDO reference.

Query 3: Genes encoding membrane transport proteins

This query targets genes whose protein products are involved in membrane transport, with emphasis on ion exchange and ATPase activity. These proteins are functionally critical for maintaining neuronal excitability and ionic homeostasis, processes frequently altered in dystonia. The query filters for the occurrence of the term “transport” in the protein label to identify genes encoding transporter proteins.

SPARQL Query: Genes encoding transport-related proteins

```
PREFIX dist_r: <https://um.es/data/dystonia/>
PREFIX dist_o: <https://dystonia_ontology.um.es/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX ro: <http://purl.obolibrary.org/obo/RO_>

SELECT DISTINCT ?geneLabel ?chromosome ?alteration ?proteinLabel
WHERE {
  ?gene a dist_o:Gene ;
        rdfs:label ?geneLabel ;
        dist_o:locatedOnChromosome ?chromosome ;
        dist_o:hasAlteration ?alteration ;
        ro:0002205 ?protein .

  ?protein rdfs:label ?proteinLabel .

  FILTER CONTAINS(LCASE(STR(?proteinLabel)), "transport")
}
ORDER BY ?geneLabel
```

Query results:

Gene	Chr	Alteration	Protein
ATP1A3	19q13.2	c.1871C>T (p.Thr624Met)	Na+/K+ transporting ATPase subunit alpha-3

Table 5: Gene encoding a membrane transporter protein relevant to neuronal function.

Query 4: Genes and their associated protein products (UniProt)

This query aims to retrieve, for each gene in the RDF dataset, the associated UniProt protein identifier and the corresponding protein name. This linkage is captured through the semantic property `ro:0002205`, which connects gene instances to their protein products in external databases. Retrieving this information is essential for integrating genomic data with proteomic knowledge and supporting downstream biological interpretation.

SPARQL Query: Genes and UniProt protein products

```
PREFIX dist_o: <https://dystonia_ontology.um.es/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX ro: <http://purl.obolibrary.org/obo/RO_>

SELECT DISTINCT ?geneLabel ?uniprotID ?proteinLabel
WHERE {
  ?gene a dist_o:Gene ;
        rdfs:label ?geneLabel ;
        ro:0002205 ?uniprotID .
  ?uniprotID rdfs:label ?proteinLabel .
}
ORDER BY ?geneLabel
```

Query results:

Gene	UniProt ID	Protein Name
ANO3	https://www.uniprot.org/uniprot/Q9BYT9	Anoctamin-3
ATP1A3	https://www.uniprot.org/uniprot/P13637	Na+/K+ TP ATPase subunit alpha-3
GCH1	https://www.uniprot.org/uniprot/P30793	GTP cyclohydrolase 1
PANK2	https://www.uniprot.org/uniprot/Q9BZ23	Pantothenate kinase 2
PRRT2	https://www.uniprot.org/uniprot/Q7Z6L0	Proline-rich transmembrane protein 2
SGCE	https://www.uniprot.org/uniprot/O43556	Epsilon-sarcoglycan
THAP1	https://www.uniprot.org/uniprot/Q9NVV9	THAP domain-containing protein 1
TOR1A	https://www.uniprot.org/uniprot/O14656	Torsin-1A

Table 6: Genes and their associated protein products as recorded in the UniProt database.

Query 5: Genes with both mutation and syndrome annotation

This query retrieves all gene instances in the RDF dataset that are annotated with both a genetic alteration and an associated clinical syndrome. Unlike previous queries that also incorporated protein-level annotations, this version focuses on the core genotype–phenotype relationship, which is often the foundation of biomedical interpretation. The query results help identify the most biologically and clinically informative gene entities, providing a concise summary of their mutational status and related syndromic outcome.

SPARQL Query: Genes with alteration and syndrome

```
PREFIX dist_o: <https://dystonia_ontology.um.es/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT DISTINCT ?geneLabel ?alteration ?syndromeLabel
WHERE {
  ?gene a dist_o:Gene ;
        rdfs:label ?geneLabel ;
        dist_o:hasAlteration ?alteration ;
        dist_o:isAssociatedWith ?syndrome .

  ?syndrome rdfs:label ?syndromeLabel .
}
ORDER BY ?geneLabel
```

Query results:

Gene	Alteration	Syndrome
ANO3	c.1409G>A (p.Gly470Glu)	Craniocervical dystonia (DYT24)
ATP1A3	c.1871C>T (p.Thr624Met)	Rapid-onset dystonia-parkinsonism
GCH1	c.317C>T (p.Thr106Ile)	Dopa-responsive dystonia
PANK2	c.1133A>G (p.Asp378Gly)	Pantothenate kinase-associated neurodegeneration
PRRT2	c.859G>A (p.Ala287Thr)	Paroxysmal kinesigenic dyskinesia (PKD)
SGCE	c.109+5G>C (splicing)	Myoclonus-dystonia
THAP1	c.97T>G (p.Cys33Gly)	Early-onset dystonia type 6 (DYT6)
TOR1A	c.907_909delGAG (p.Glu303del)	DYT1 early-onset generalized dystonia

Table 7: Genes annotated with both a genetic alteration and an associated clinical syndrome.

4 Conclusions

This project highlights the practical applicability and semantic expressiveness of Semantic Web technologies for modeling domain-specific biomedical knowledge. By focusing on dystonia-associated genes and their clinical manifestations, the work successfully demonstrates how ontological modeling and RDF encoding can transform curated biomedical data into interoperable, machine-readable knowledge graphs.

1. A custom ontology was designed and implemented in Turtle format to formally represent key biological entities using well-defined classes and properties aligned with external ontologies such as SO, DOID, RO, and SIO.

2. The RDF dataset was accurately converted into RDF/XML format to ensure compatibility with triple stores like Blazegraph. Syntax validation confirmed full compliance with RDF standards, ensuring semantic integrity and technical interoperability.

3. The Blazegraph Workbench facilitated efficient query development and inspection of the dataset. Complementarily, the implementation of SPARQL queries in R illustrated the potential for programmatic access and reproducible semantic querying within FAIR-compliant data workflows.

4. All SPARQL queries produced biologically coherent results, supporting mutation patterns ,genotype–phenotype associations, and protein-level annotations. This confirmed the consistency and queryability of the knowledge graph.

Future work could expand this model by incorporating additional gene–phenotype relationships, regulatory elements, and transcriptomic data, thus enabling multi-omic integration and more advanced reasoning over complex neurogenetic disorders.

5 Availability of Resources

All relevant files generated in the context of this project have been made publicly available in the following GitHub repository: https://github.com/carlos-um/ESD. This

repository includes the ontology in Turtle format (.ttl), used for data modeling and initial editing; the RDF version in .rdf format, used for uploading the dataset to Blazegraph and performing SPARQL queries; and the R script used to execute queries against the Blazegraph endpoint.

These resources allow for full reproducibility of the data publication process and facilitate further exploration of the semantic dataset.

References

[1] Bettina Balint et al. "Dystonia". In: *Nature Reviews Disease Primers* 4.1 (Sept. 2018). ISSN: 2056-676X. DOI: 10.1038/s41572-018-0023-6. URL: http://dx.doi.org/10.1038/s41572-018-0023-6.

[2] Amit Batla. "Dystonia: A review". In: *Neurology India* 66.7 (2018), p. 48. ISSN: 0028-3886. DOI: 10.4103/0028-3886.226439. URL: http://dx.doi.org/10.4103/0028-3886.226439.

[3] H.A. Jinnah and Yan V. Sun. "Dystonia genes and their biological pathways". In: *Neurobiology of Disease* 129 (Sept. 2019), pp. 159–168. ISSN: 0969-9961. DOI: 10.1016/j.nbd.2019.05.014. URL: http://dx.doi.org/10.1016/j.nbd.2019.05.014.

[4] Jun Ma et al. "Targeted gene capture sequencing in diagnosis of dystonia patients". In: *Journal of the Neurological Sciences* 390 (July 2018), pp. 36–41. ISSN: 0022-510X. DOI: 10.1016/j.jns.2018.04.005. URL: http://dx.doi.org/10.1016/j.jns.2018.04.005.

Appendix A: RDF dataset in Turtle format

dystonia.ttl

```
############ PREFIXES ############

@prefix dist_r: <https://um.es/data/dystonia/> .       # Instances
@prefix dist_o: <https://dystonia_ontology.um.es/> .   # Ontology terms
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix ncbi: <https://www.ncbi.nlm.nih.gov/gene/> .
@prefix uniprot: <https://www.uniprot.org/uniprot/> .
@prefix omim: <https://omim.org/entry/> .
@prefix mondo: <http://purl.obolibrary.org/obo/MONDO_> .
@prefix doid: <http://purl.obolibrary.org/obo/DOID_> .
@prefix ro: <http://purl.obolibrary.org/obo/RO_> .
@prefix sio: <http://semanticscience.org/resource/> .

############ CLASSES ############

dist_o:Gene rdf:type owl:Class ;
    rdfs:label "Gene" ;
    owl:equivalentClass <http://purl.obolibrary.org/obo/SO_0000704> .

dist_o:Syndrome rdf:type owl:Class ;
    rdfs:label "Syndrome" ;
    owl:equivalentClass <http://purl.obolibrary.org/obo/DOID_225> .

############ PROPERTIES ############

dist_o:hasAlteration rdf:type owl:DatatypeProperty ;
    rdfs:label "has alteration" ;
    rdfs:comment "Describes a genetic mutation or variation in the gene." ;
    owl:equivalentProperty <http://semanticscience.org/resource/SIO_001114> .

dist_o:isAssociatedWith rdf:type owl:ObjectProperty ;
    rdfs:label "is associated with" ;
    rdfs:comment "Relates a gene to a clinical syndrome or disease." ;
    owl:equivalentProperty <http://semanticscience.org/resource/SIO_000628> .

dist_o:locatedOnChromosome rdf:type owl:DatatypeProperty ;
    rdfs:label "located on chromosome" ;
    rdfs:comment "Specifies the chromosomal location of a gene." ;
    rdfs:domain dist_o:Gene ;
    rdfs:range xsd:string .

ro:0002205 rdf:type owl:ObjectProperty ;
    rdfs:label "encodes" ;
    rdfs:comment "Relates a gene to the protein it encodes." .
```

dystonia.ttl

```
############ INSTANCES: GENES + SYNDROMES ############

dist_r:TOR1A rdf:type dist_o:Gene ;
    rdfs:label "TOR1A" ;
    owl:sameAs ncbi:1861 ;
    ro:0002205 uniprot:O14656 ;
    dist_o:locatedOnChromosome "9q34.11"^^xsd:string ;
    dist_o:hasAlteration "c.907_909delGAG (p.Glu303del)" ;
    rdfs:seeAlso omim:128100 ;
    dist_o:isAssociatedWith dist_r:DYT1 .

uniprot:O14656 rdfs:label "Torsin-1A" .

dist_r:DYT1 rdf:type dist_o:Syndrome ;
    rdfs:label "DYT1 early-onset generalized dystonia" ;
    owl:sameAs mondo:0007492 .

dist_r:PANK2 rdf:type dist_o:Gene ;
    rdfs:label "PANK2" ;
    owl:sameAs ncbi:80025 ;
    ro:0002205 uniprot:Q9BZ23 ;
    dist_o:locatedOnChromosome "20p13"^^xsd:string ;
    dist_o:hasAlteration "c.1133A>G (p.Asp378Gly)" ;
    rdfs:seeAlso omim:234200 ;
    dist_o:isAssociatedWith dist_r:PKAN .

uniprot:Q9BZ23 rdfs:label "Pantothenate kinase 2" .

dist_r:PKAN rdf:type dist_o:Syndrome ;
    rdfs:label "Pantothenate kinase-associated neurodegeneration (PKAN)" ;
    owl:sameAs mondo:0009319 .

dist_r:ATP1A3 rdf:type dist_o:Gene ;
    rdfs:label "ATP1A3" ;
    owl:sameAs ncbi:478 ;
    ro:0002205 uniprot:P13637 ;
    dist_o:locatedOnChromosome "19q13.2"^^xsd:string ;
    dist_o:hasAlteration "c.1871C>T (p.Thr624Met)" ;
    rdfs:seeAlso omim:128235 ;
    dist_o:isAssociatedWith dist_r:RapidDystonia .

uniprot:P13637 rdfs:label "Sodium/potassium-transporting ATPase subunit alpha-3" .

dist_r:RapidDystonia rdf:type dist_o:Syndrome ;
    rdfs:label "Rapid-onset dystonia-parkinsonism" ;
    owl:sameAs mondo:0007496 .
```

dystonia.ttl

```
dist_r:PRRT2 rdf:type dist_o:Gene ;
    rdfs:label "PRRT2" ;
    owl:sameAs ncbi:116512 ;
    ro:0002205 uniprot:Q7Z6L0 ;
    dist_o:locatedOnChromosome "16p11.2"^^xsd:string ;
    dist_o:hasAlteration "c.859G>A (p.Ala287Thr)" ;
    rdfs:seeAlso omim:128200 ;
    dist_o:isAssociatedWith dist_r:PKD .

uniprot:Q7Z6L0 rdfs:label "Proline-rich transmembrane protein 2"

dist_r:PKD rdf:type dist_o:Syndrome ;
    rdfs:label "Paroxysmal kinesigenic dyskinesia (PKD)" ;
    owl:sameAs mondo:0044202 .

dist_r:GCH1 rdf:type dist_o:Gene ;
    rdfs:label "GCH1" ;
    owl:sameAs ncbi:2643 ;
    ro:0002205 uniprot:P30793 ;
    dist_o:locatedOnChromosome "14q22.2"^^xsd:string ;
    dist_o:hasAlteration "c.317C>T (p.Thr106Ile)" ;
    rdfs:seeAlso omim:128230 ;
    dist_o:isAssociatedWith dist_r:DRD .

uniprot:P30793 rdfs:label "GTP cyclohydrolase 1" .

dist_r:DRD rdf:type dist_o:Syndrome ;
    rdfs:label "Dopa-responsive dystonia" ;
    owl:sameAs mondo:0016812 .

dist_r:THAP1 rdf:type dist_o:Gene ;
    rdfs:label "THAP1" ;
    owl:sameAs ncbi:55145 ;
    ro:0002205 uniprot:Q9NVV9 ;
    dist_o:locatedOnChromosome "8p11.21"^^xsd:string ;
    dist_o:hasAlteration "c.97T>G (p.Cys33Gly)" ;
    rdfs:seeAlso omim:605909 ;
    dist_o:isAssociatedWith dist_r:DYT6 .

uniprot:Q9NVV9 rdfs:label "THAP domain-containing protein 1" .

dist_r:DYT6 rdf:type dist_o:Syndrome ;
    rdfs:label "Early-onset dystonia type 6 (DYT6)" ;
    owl:sameAs mondo:0100016 .
```

dystonia.ttl

```
dist_r:SGCE rdf:type dist_o:Gene ;
    rdfs:label "SGCE" ;
    owl:sameAs ncbi:8910 ;
    ro:0002205 uniprot:O43556 ;
    dist_o:locatedOnChromosome "7q21.3"^^xsd:string ;
    dist_o:hasAlteration "c.109+5G>C (splicing)" ;
    rdfs:seeAlso omim:159900 ;
    dist_o:isAssociatedWith dist_r:Myoclonus .

uniprot:O43556 rdfs:label "Epsilon-sarcoglycan" .

dist_r:Myoclonus rdf:type dist_o:Syndrome ;
    rdfs:label "Myoclonus-dystonia" ;
    owl:sameAs mondo:0000903 .

dist_r:ANO3 rdf:type dist_o:Gene ;
    rdfs:label "ANO3" ;
    owl:sameAs ncbi:283570 ;
    ro:0002205 uniprot:Q9BYT9 ;
    dist_o:locatedOnChromosome "11p14.3-p14.2"^^xsd:string ;
    dist_o:hasAlteration "c.1409G>A (p.Gly470Glu)" ;
    rdfs:seeAlso omim:615034 ;
    dist_o:isAssociatedWith dist_r:DYT24 .

uniprot:Q9BYT9 rdfs:label "Anoctamin-3" .

dist_r:DYT24 rdf:type dist_o:Syndrome ;
    rdfs:label "Craniocervical dystonia (DYT24)" ;
    owl:sameAs mondo:0011886 .
```

Appendix B: SPARQL queries implemented in R

sparql_queries.R

```
library(SPARQL)

endpoint <- "http://dayhoff.inf.um.es:3050/blazegraph/namespace/dystonia/sparql"

# ---------------------------------------------------------------
# Query 1: Genes with splicing-related alterations
# ---------------------------------------------------------------
query1 <- "
PREFIX dist_r: <https://um.es/data/dystonia/>
PREFIX dist_o: <https://dystonia_ontology.um.es/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT DISTINCT ?geneLabel ?chromosome ?alteration ?syndromeLabel
WHERE {
  ?gene a dist_o:Gene ;
        rdfs:label ?geneLabel ;
        dist_o:hasAlteration ?alteration ;
        dist_o:locatedOnChromosome ?chromosome ;
        dist_o:isAssociatedWith ?syndrome .
  ?syndrome rdfs:label ?syndromeLabel .
  FILTER CONTAINS(LCASE(STR(?alteration)), \"splicing\")
}
ORDER BY ?geneLabel
"
cat("Query 1: Genes with splicing-related alterations\n")
result1 <- SPARQL(endpoint, query1)
View(result1$results)

# ---------------------------------------------------------------
# Query 2: Clinical syndromes associated with the gene ANO3
# ---------------------------------------------------------------
query2 <- "
PREFIX dist_r: <https://um.es/data/dystonia/>
PREFIX dist_o: <https://dystonia_ontology.um.es/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>

SELECT DISTINCT ?geneLabel ?chromosome ?syndromeLabel ?mondoID
WHERE {
  dist_r:ANO3 rdfs:label ?geneLabel ;
              dist_o:locatedOnChromosome ?chromosome ;
              dist_o:isAssociatedWith ?syndrome .
  ?syndrome rdfs:label ?syndromeLabel ;
            owl:sameAs ?mondoID .
}
"
cat("Query 2: Syndromes associated with ANO3\n")
result2 <- SPARQL(endpoint, query2)
View(result2$results)
```

sparql_queries.R

```
# ---------------------------------------------------------------
# Query 3: Genes encoding membrane transport proteins
# ---------------------------------------------------------------
query3 <- "
PREFIX dist_r: <https://um.es/data/dystonia/>
PREFIX dist_o: <https://dystonia_ontology.um.es/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX ro: <http://purl.obolibrary.org/obo/RO_>

SELECT DISTINCT ?geneLabel ?chromosome ?alteration ?proteinLabel
WHERE {
  ?gene a dist_o:Gene ;
        rdfs:label ?geneLabel ;
        dist_o:locatedOnChromosome ?chromosome ;
        dist_o:hasAlteration ?alteration ;
        ro:0002205 ?protein .
  ?protein rdfs:label ?proteinLabel .
  FILTER CONTAINS(LCASE(STR(?proteinLabel)), \"transport\")
}
ORDER BY ?geneLabel
"
cat("Query 3: Genes encoding transport proteins\n")
result3 <- SPARQL(endpoint, query3)
View(result3$results)

# ---------------------------------------------------------------
# Query 4: Genes and their associated UniProt proteins
# ---------------------------------------------------------------
query4 <- "
PREFIX dist_o: <https://dystonia_ontology.um.es/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX ro: <http://purl.obolibrary.org/obo/RO_>

SELECT DISTINCT ?geneLabel ?uniprotID ?proteinLabel
WHERE {
  ?gene a dist_o:Gene ;
        rdfs:label ?geneLabel ;
        ro:0002205 ?uniprotID .
  ?uniprotID rdfs:label ?proteinLabel .
}
ORDER BY ?geneLabel
"
cat("Query 4: Genes and associated UniProt proteins\n")
result4 <- SPARQL(endpoint, query4)
View(result4$results)
```

sparql_queries.R

```
# ---------------------------------------------------------------
# Query 5: Genes with both mutation and syndrome annotation
# ---------------------------------------------------------------
query5 <- "
PREFIX dist_o: <https://dystonia_ontology.um.es/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT DISTINCT ?geneLabel ?alteration ?syndromeLabel
WHERE {
  ?gene a dist_o:Gene ;
        rdfs:label ?geneLabel ;
        dist_o:hasAlteration ?alteration ;
        dist_o:isAssociatedWith ?syndrome .
  ?syndrome rdfs:label ?syndromeLabel .
}
ORDER BY ?geneLabel
"
cat("Query 5: Genes with mutation and syndrome annotation\n")
result5 <- SPARQL(endpoint, query5)
View(result5$results)
```