DisGeNET RDF: a gene-disease association Linked Open Data resource

Núria Queralt-Rosinach and Laura I. Furlong

Research Programme on Biomedical Informatics (GRIB),
Hospital del Mar Medical Research Institute (IMIM),
Pompeu Fabra University (UPF),
C/ Dr. Aiguader 88, 08003 Barcelona, Spain
http://ibi.imim.es

Abstract. *We present the first RDF representation of DisGeNET, a gene-disease database designed to integrate the current knowledge of human diseases. DisGeNET RDF data introduces a harmonized and semantically enriched description of the gene-disease association concept into the Semantic Web (SW) by means of the DisGeNET ontology. The centric view on gene-disease associations allows to find links between diseases and genes easily and with added semantic content. The RDF representation follows the Linked Data (LD) principles that provides new opportunities for data integration, querying and crossing DisGeNET data with other external RDF datasets. The RDF version of DisGeNET has been developed in the context of the Open PHACTS project to provide disease relevant information to the knowledge base on pharmacological data.*

Keywords: gene-disease association, RDF, ontology, linked data

1 Introduction

DisGeNET RDF is presented as a new resource in the Linked Open Data space to promote the discovery of key knowledge in the understanding of the molecular mechanisms underlying a disease or an adverse drug event. DisGeNET [1, 2] is a relational database that integrates gene-disease associations from curated databases and the literature, and additional information such as pathways or SNPs. DisGeNET has been mapped to RDF using the most advanced Semantic Web technologies such as OWL and SPARQL. In this workshop, we are introducing DisGeNET as a new RDF resource in the LD space, the methodology used for its conversion, the specific ontology developed to model the gene-disease association concept and, finally, some potential applications of the resource.

2 Methodology

2.1 The RDF Schema, ontologies and identifiers

The RDF version of DisGeNET is represented as a set of triples around the gene-disease association concept. Information such as genomic variation or sci-

entific publications supporting the association are related to this main concept. Data is organized in a hierarchical manner around five concepts: gene, disease, pathway, disease class, and gene-disease association as the parent concept (see DisGeNET RDF web interface `http://rdf.imim.es/DisGeNET.html` for details on the RDF schema). The 'RDF-ization' has been done using RDFS and OWL languages, common ontologies and vocabularies, and following the Linked Data principles (`http://linkeddata.org/`). To identify resources in DisGeNET, URIs established by the identifiers.org effort were selected whenever possible to support its initiative to foster a unified use of URIs by the SW community [3]. DisGeNET data is open and is linked out to other linked data resources such as Linked Life Data (`http://linkedlifedata.com/`) or Bio2RDF projects [4].

2.2 DisGeNET gene-disease association ontology

The DisGeNET gene-disease association ontology harmonizes the semantic description of the different types of associations between genes and diseases. The ontology provides foundational support for the DisGeNET database. This ontology was integrated in the increasingly used Semanticscience Integrated Ontology (SIO) [5], which is an ontology meant to be adopted to describe basic scientific semantics and ensure correct concept mapping among other more specific ontologies. The DisGeNET ontology can be accessed at `http://ibi.imim.es/DisGeNET-Dev/ontologies/GeneDiseaseAssociation.owl`.

2.3 Provenance description

It is considered good practice to provide provenance information to an RDF resource. The provenance description of the DisGeNET RDF dataset declares the database from which it is derived, the development date, the current version and updates, the software used for its development, the license information, the SPARQL endpoint location, the number of triples, etc... The provenance description of the original database and each primary source is also tracked. This information is provided using the Vocabulary of Interlinked Datasets (VoID) [6].

2.4 Data processing

To map the relational database content into RDF triples, we used the D2RQ platform (`http://d2rq.org/`). The dump files and the VoID description of DisGeNET are loaded into the OpenLink Virtuoso RDF Quad Store [7]. An SPARQL endpoint hosted in the Virtuoso server has been implemented as the primary interface to access the RDF data (for access to RDF data see the DisGeNET RDF web interface). Validation of data was done with Protegé platform (`http://protege.standford.edu`) and our SPARQL endpoint.

3 Results: Integration across resources

We present the potential of DisGeNET Linked Data with different uses cases. The first use case is aimed at answering the following question: *give me all the gene-disease associations from the Comparative Toxicogenomics Database (CTD) that have information on sequence variation and provenance.* This question can be answered exploiting the information contained in DisGeNET, selecting the associations provided by CTD and that have annotations on genomic variation and literature provenance. This is translated in the following SPARQL query:

```
SELECT DISTINCT ?gdassocIRI ?dName ?gName ?snpLabel ?PMIDLabel
WHERE
{
  ?gdassocIRI sio:SIO_000628 ?disease,?gene .
  ?disease rdf:type ncit:C7057 .
  ?disease foaf:name ?dName .
  ?gene rdf:type ncit:C16612 .
  ?gene sio:SIO_000205 ?genenameURI .
  ?genenameURI rdfs:label ?gName .
  ?gdassocIRI sio:SIO_000001 ?snp .
  ?snp rdf:type ncit:C18279 .
  ?snp rdfs:label ?snpLabel .
  ?gdassocIRI sio:SIO_000253 ?source .
  FILTER regex(?source,"ctd")
  ?gdassocIRI sio:SIO_000772 ?PubMedID .
  ?PubMedID rdfs:label ?PMIDLabel .
}
```

The inclusion of DisGeNET into the 'Web of linked data' using the most advanced SW technologies brings the opportunity to integrate our gene-disease data with other disparate data sources spread over the Web by performing federated queries. The second use case is aimed at answering the following question: *give me all the gene-disease associations in which the association is linked to changes in the expression of the gene, there is expression information, and sequence variation linked to the disease.* This question can be answered by querying and integrating data from DisGeNET and GXA (`http://www.ebi.ac.uk/gxa/`) databases. More specifically, we ask for the DisGeNET gene-disease associations labelled as 'AlteredExpression' and their related SNPs, and the GXA expression values. See the SPARQL query at the DisGeNET RDF web interface. DisGeNET SPARQL endpoint supports the syntax and semantics of SPARQL 1.1 for executing queries distributed over different SPARQL endpoints. SPARQL queries such these are aimed to be included in Bioqueries which is a portal aimed at gathering SPARQL queries [8].

As the last use case, DisGeNET RDF has been implemented in the Open Pharmacological Space (OPS) discovery platform, which is a SW platform devel-

oped under the Innovative Medicines Initiative (IMI; http://www.imi.europa.eu) funded Open PHACTS project. Remarkably, the integration of DisGeNET in OPS is essential to answer important research questions such as which compounds could effectively inhibit targets involved in a key pathway for the development of a disease. Aimed at exploring and querying DisGeNET data across the linked data in the platform, APIs are currently under development (see the API website for up to date information at http://dev.openphacts.org). This is expected to be fully operative in the upcoming OPS 1.5 release.

4 Conclusions

We present DisGeNET RDF, a new Linked Data dataset that provides gene-disease association data to answer relevant scientific pharmacological complex questions. Importantly, DisGeNET has been implemented in the pharmacological LD discovery platform developed within the Open PHACTS project.

Acknowledgments. The research leading to these results has received support from the IMI Joint Undertaking under grant agreement n° 115191, Open PHACTS, resources of which are composed of financial contribution from the EU FP7 (FP7/2007-2013) and EFPIA companies' in kind contribution; and the Instituto de Salud Carlos III FEDER (CP10/005249). The Research Programme on Biomedical Informatics (GRIB) is a node of the Spanish National Institute of Bioinformatics (INB).

References

1. Bauer-Mehren, A., Rautschka, M., Sanz, F., Furlong, L.I.: DisGeNET: a Cytoscape Plugin to Visualize, Integrate, Search and Analyze Gene-Disease Networks. BMC Bioinformatics. 26, 2924–2926 (2010)
2. Bauer-Mehren, A., Bundschus, M., Rautschka, M., Mayer, M.A., Sanz, F., Furlong, L.I.: Gene-Disease Network Analysis Reveals Functional Modules in Mendelian, Complex and Environmental Diseases. PLOS One. 6, e20284 (2011)
3. Juty, N., Le Nov, N.: Identifiers.org and MIRIAM Registry: Community Resources to Provide Persistent Identification. Nucleic Acids Res. 40, D580–D586 (2012)
4. Belleau, F.: Bio2RDF: towards a Mashup to Build Bioinformatics Knowledge Systems. J. Biomed. Inform. 41, 706–716 (2008)
5. Dumontier et al., The Semanticscience Integrated Ontology (SIO) for Biomedical Research and Knowledge Discovery. (2013) (*Submitted*)
6. Describing Linked Datasets with the VoID Vocabulary. W3C Interest Group Note, 3 March 2011. http://www.w3.org/TR/void/
7. Erling, O. and Mikhailov, I. Virtuoso: RDF support in a native RDBMS, page 501 (2010)
8. Godoy, M.J., Lopez-Camacho, E., Navas-Delgado, I. and Aldana-Montes, J.F. Sharing and executing linked data queries in a collaborative environment. Bioinformatics, pages 1-8 (2013)