DisGeNET: from MySQL to Nanopublication, Modelling Gene-Disease Associations for the Semantic Web

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

Abstract. DisGeNET is a relational gene-disease database that has been converted to RDF in order to make gene-disease association data available for Semantic Web projects such as Open PHACTS. In this paper, the conversion of DisGeNET from MySQL to RDF and their modelization to the nanopublication data format is presented, and we discuss the challenges encountered throughout the process.

Keywords: Semantic Web, gene-disease association, relational database, RDF, ontology, nanopublication.

1 Introduction

The ideal data infrastructure for a pharmaceutical researcher is one that makes it easy to carefully assemble, overlay and search across heterogeneous data sources in order to extract knowledge to solve drug discovery complex questions. The RDF Semantic Web (SW) technology has gained significant presence in the Life Sciences to connect the various databases in this field. The Open PHACTS (Open Pharmacological Concept Triple Store) is a project funded by a European grant from the Innovative Medicines Initiative (IMI; `http://www.imi.europa.eu`) that aims to integrate distributed heterogenous data sources in a SW approach, developing an open source, open standards and open access innovation platform, the Open Pharmacological Space (OPS). The project intends to reach this goal by using the Linked Data approach (`http://linkeddata.org`) and managing the data in an RDF triple store. This semantically enriched and fully interoperable platform currently contains the relationships between compound-target-pathway concepts and, consequently, it delivers information on small molecules and their pharmacological profiles as well as on biological targets and pathways. But, it is necessary the addition of known gene-disease associations to answer important research questions that cannot be addressed with the existing OPS, such as which compounds could effectively inhibit targets involved in a key pathway for the development of a disease, to explore potential toxic interactions, or drug repositioning opportunities in new therapeutic areas.

In our lab a relational database called DisGeNET [1, 2] was created in order to contain the current knowledge of human genetic diseases including mendelian, complex and environmental diseases. DisGeNET is a comprehensive gene-disease database that integrates gene-disease associations stored in UniProt, CTD™, GAD, MGD databases and text-mining derived associations from the literature-derived human gene-disease network (LHGDN) database. In addition, gene-pathway information from Reactome and SNPs associated to gene-disease relationships is provided in order to have a more complete picture of the biological processes underlying a disorder and the correlation of specific genomic variants with disease predisposition. The integration is performed by means of gene-disease vocabulary mapping and by using a new gene-disease association ontology. Since source databases use two different disease vocabularies (MIM and MeSH terms), a vocabulary mapping is done by means of the UMLS® Metathesaurus® concept structure. Therefore, DisGeNET in RDF could be implemented in OPS enabling the inclusion of disease-gene-pathway concepts in the platform and to integrate its data with OPS compound/drug data. For this reason, the DisGeNET MySQL database has been converted into the RDF data model. Moreover, as the Open PHACTS project is co-developing and exploiting the nanopublication format, which allows individual data to be publishable, cited and attributed in a RDF-based approach, we are adapting our RDF DisGeNET data to the OPS nanopublication model according to the latest Open PHACTS guidelines, since our data can benefit from its citability and publishable features.

In this paper, we present DisGeNET as a new RDF gene-disease association database, the methodology used for the MySQL-RDF conversion, the new ontology developed to model the gene-disease association concept, the nanopublication data model, and, finally we discuss some of the challenges encountered throughout the process.

2 Results

To convert a relational database to a RDF we first identified the lists of concepts and relations from our data. Once this was done, a RDF data model schema that represents the knowledge stored in our database was created. Our RDF data model captures the central role that gene-disease associations play in our database to comprise the whole spectrum of human diseases with a genetic origin. In an RDF data representation model, the information from different data sources is semantically connected to each other using existing commonly shared ontologies. Then, we explored the existing ontologies via services such as BioPortal in order to find matching entries in those existing ontologies for each of our concepts and relations. RDF properties were mapped onto a limited set of external ontologies and vocabularies that include the SemanticScience Integrated Ontology (SIO) for general science, NCI Thesaurus for biomedical terms, and Dublin Core to encode license information. We also used common vocabularies such as rdf:, rdfs:, and owl:. Resources, i.e. objects and subjects in RDF triples,

were identified by dereferenceable Internationalized Resource Identifiers (IRIs) built upon DisGeNET IDs, which are IDs of other data collections. The providers of these IRIs are the new Identifiers.org service (`http://identifiers.org`) + the MIRIAM Registry [4], and the Bio2RDF project [3]. Nevertheless, as the disease concept in our database is identified by the Unified Medical Language System® Concept Unique Identifier (UMLS® CUI), we decided to use the Linked Life Data (`http://linkedlifedata.com`) provider instead of the Human Disease Ontology (the later also integrated in identifiers.org) because is directly based on the UMLS® CUI.

Common ontologies have been used whenever possible, but in the case of describing DisGeNET gene-disease association resources, new semantic terms had to be created because no similar viable terms exist. Therefore, the RDF conversion of DisGeNET is accompanied by a new gene-disease association ontology developed in our lab for a correct semantic integration of gene-disease association data from diverse data sources. For generating RDF triples we used the D2RQ platform (`http://d2rq.org`) and the RDF/Turtle language. Validation of data was done with Protegé platform (`http://protege.standford.edu`).

3 Discussion

It is well known that the namespace of biomedicine is messy and ambiguous and lacks universal standards unlike other disciplines. But, this problem is not exclusive to the identification of resources; many synonyms exist on the Web for key concept classes such as genes, proteins, genetic variations and diseases. For this reason, the most difficult part in the RDF conversion of DisGeNET was to find proper IRIs for properties and resources but, also, adequate namespaces for semantic types of concept classes. An exhaustive search for ontologies was made since we tried to choose those ontologies that fit best with the meaning of our concepts/properties and that are commonly used by the scientific community but, also, in the Open PHACTS project. An important problem not yet solved is the use of valid IRIs to describe the RDF nodes for gene-disease association concept. This is a major task as it is required that IRIs are dereferenceable, i.e. identifiers for which is possible to get information about the referenced resource on the Web. There are some possible solutions such as registering each instance of the concept to the MIRIAM Registry. Another issue still not adressed is to use a valid IRI pattern to identify disease MeSH hierarchy classes, as MeSH does not have an IRI pattern available. The nanopublication format raises another example of this IRI problem as each named graph of a nanopublication and the entire nanopublication unit itself needs an IRI pattern schema. Another problem is the proper tracking of the several modifications that a nanopublication may have due to updating/curation processes over time. We are currently tackling all these issues.

Regarding licensing, DisGeNET is distributed under the GNU GPL 3.0 license. This means that we have made data available as open data. License incompatibilities are omnipresent in open source data. This opens the question about using IRIs from databases with no open data licenses in our RDF database. Is the use of IRIs subject to licensing?

4 Conclusions

To sum up, we have carried out the RDF conversion of DisGeNET, a derivative relational database that integrates data connected with the gene-disease association concept from several public data sources (curated databases and literature), in order to integrate it in the OPS platform of the Open PHACTS project. Importantly, DisGeNET could provide significant data to answer relevant scientific pharmacological complex questions thanks to the introduction of the disease concept into OPS and its relationship with genotype. Specifically, the RDF version of DisGeNET is a set of triples that include information around gene-disease associations, such as SNPs related in the bibliography to predispositione to the disease, and the pathways where genes are known to be involved. In the main, this conversion has been done according to the Linked Data principles, open access and interoperability of the data. Currently, we are tackling the modelization of the RDF triples to the nanopublication format because this model could allow both to better adapt to the features of OPS platform and to benefit from their own advantages as the citability. In the future, the implementation of a SPARQL endpoint to provide open access to the information and to query RDF DisGeNET data will be evaluated.

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. Belleau, F.: Bio2RDF: towards a Mashup to Build Bioinformatics Knowledge Systems. J. Biomed. Inform. 41, 706–716 (2008)
4. Juty, N., Le Nov, N.: Identifiers.org and MIRIAM Registry: Community Resources to Provide Persistent Identification. Nucleic Acids Res. 40, D580–D586 (2012)