Semantic Rules on Drug Discovery Data

Sashikiran Challa, David Wild, Ying Ding, and Qian Zhu

Indiana University, 901 E 10th Street
Bloomington, Indiana, USA
{schalla,djwild,dingying,qianzhu}@indiana.edu
http://www.informatics.indiana.edu

Abstract. Aggregating and presenting a wide variety of information pertinent to the biological and pharmacological effects of chemical compounds will be a critical part of 21st century drug discovery. However there is currently a lack of tools for effectively integrating and aggregating information about chemical compound. In this paper we tackle this problem using Semantic Web Technologies, particularly OWL ontologies,compound-centric RDF networks, and RDF inference to detect relationships between compounds and biological affects, genes,and diseases, and to present information to a user clustered by disease area.

Key words: Ontology, RDF data model, Jena, reasoner

1 Introduction

In the field of drug discovery there are many repositories of information of many different types.Many are publicly available. At Indiana University, we recently developed an infrastructure to make a wide variety of these sources available as web services [4], including database services (including PubChem Compound, Pubchem BioAssay, and Drug Bank) computation services (particularly predictive models that predict biological properties of compounds), and literature searches which can identify the compounds and ontological terms that are contained in scholarly journal articles. Aggregating information about a compound from several different web services is being achieved using a tool called WENDI (Web Engine for Non-obvious Drug Information). We have now taken the next step which is to not only integrate information but also make some new inferences and establish relationships among the data aggregated using OWL, RDF and reasoning engines.

Drug discovery data used here was collected from the WENDI aggregate web service. WENDI takes a single compound as a query and then aggregates comprehensive information about a compound from web services that represent several diverse sources (including predictive models,chemical compound databases, and journal articles). The aggregate information was obtained as an XML document from which active compounds, journal articles information was extracted using minidom package in Python Scripting language. This information was converted into RDF triples[2] using Python according to Wendi Ontology (created

in house using Protege4.0). The Ontology has Chemical Compound, BioAssay, Journal Article as Classes;isSimilarTo, isAssociatedWith, isContainedIn as the Object Properties; hasPubMedID,hasTitle as Data Type Properties.Here are some of the triples in Turtle format generated based on the Wendi Ontology. These triples mean that a compound with ID 15940175 is a type of Chemical Compound and that it is associated with a Bio Assay with an ID 1004.

```
WO:cid15940175 rdf:type owl:ChemicalCompound;
               WO:isAssociatedWith WO:aid1004.
```

2 Framing the Rules

RDF triples thus generated based on Wendi Ontology were loaded into Ont Model class in Jena, a java framework for building semantic web applications. Then rules were written as shown below.

```
[rule1:(?querycmpd WO:isSimilarTo ?cid)
(?cid  WO:isContainedIn ?journal)
->(?qerycmpd  WO:mightBeContainedIn ?journal)];
[rule2:(?querycmpd WO:isSimilarTo ?cid)
(?cid WO:isAssociatedWith ?aid)
->(?querycmpd WO:mightBeAssociatedWith ?aid)];
```

The rule1 means that if there is a triple with 'isSimilarTo' as property, query compound as subject and say a compound C as object and if there is a triple with 'isContainedIn' as property, compound C as subject, JournalId as object, then infer a relationship between query compound and JournalId by creating a triple with 'mightBeContainedIn' as property, query compound as subject, JournalId as object. Similar is rule2 which infers new property 'mightBeAssociatedWith' between query compound and Bio Assay ID. Rules can be extended to include diseases, gene names. The rules were parsed using the Generic rule reasoner belonging to Reasoner class in Jena. On these additionally generated triples, SPARQL queries were written to output the information about the query compound, Assay ids, Journal article titles, and their Pubmed ids.

An example of the kinds of inference we can produce: If a compound A was to found to be similar to compound B and if compound B was found to be active in a Bio Assay C, then an inference that Compound A might be active in Bio Assay C is one such inference. And say if a compound A was found to be similar to compound B and if compound B was found to be contained in a Journal C, then an inference that compound A's relevant information could be contained in Journal C is another inference achieved.

3 Conclusion

We are now able to infer relationships between compounds, genes and diseases based on RDF chains which are interpretable by a medicinal chemist. Thus we

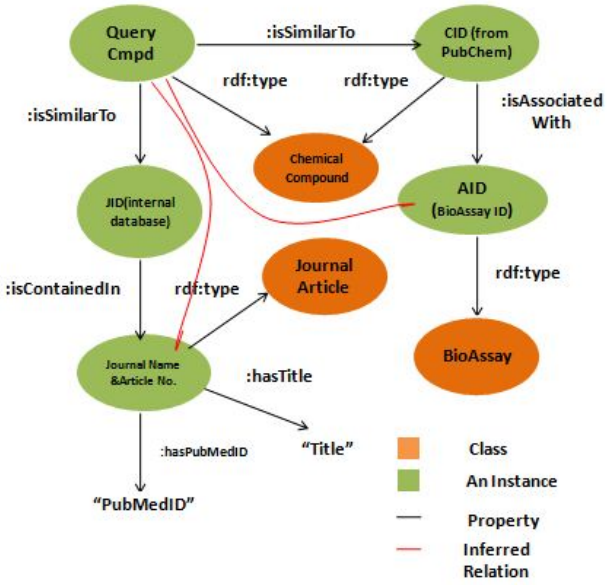

Fig. 1. RDF graph generated on Drug Discovery Data

can not only associate a compound with a disease, say, but also present the evidence for this association. Thus the work described here helps us to realize the potential of having the data as RDF data model triples based on an Ontology and the potential of rules in data integration and knowldege discovery.

4 References

References

1. Stephens S.:Enabling Semantic Web Inferencing with Oracle Technology: Applications in Life Sciences. Lecture Notes in Computer Sciences. 3791, pp. 8-16 (2005)
2. Wang, X., Gorlitsky Robert, Almeida,J.S.: From XML to RDF: how semantic web technologies will change the design of 'omic' standards.Nature Biotechnology, vol.23, 9, pp. 1099-1103. (May 2005)
3. Hugo,Lam,YK., Marenco, Luis. Clark, Tim. Gao, Yong. Kinoshita, June. Shepherd, Gordon. Miller, Perry. Wu, Elizabeth. Wong, T. Gwendolyn. Liu, Nian. Crasto. Chiquito. Morse, Thomas. Stepehen, Susie. Cheung,K: AlzPharm: integration of neurodegeneration data using RDF.BMC Bioinformatics.8 (2007)
4. Dong, X., Gilbert, K.E., Guha, R., Heiland, R., Kim, J., Pierce, M.E., Fox, G.C. and Wild, D.J. Web service infrastructure for chemoinformatics, Journal of Chemical Information and Modeling,47(4) pp 1303-1307. 2007