Representing Text Mining Results for Structured Pharmacological Queries

Carina Haupt[1], Paul Groth[2] and Marc Zimmermann[1],

[1] Bonn-Aachen International Center for Information Technology B-IT, Dahlmannstraße 2, 53113 Bonn, Germany
{hauptc, mzimmerm}@informatik.uni-bonn.de

[2] Department of Computer Science, VU University Amsterdam, De Boelelaan 1081a, 1081 HV Amsterdam, Netherlands
p.t.groth@vu.nl

Abstract. Several approaches integrating life science data using Semantic Web technologies have been described in the literature. However, these approaches have largely ignored the vast amount of content only available within the scientific literature. In this article, we present an RDF schema for text mining results that enables queries in SPARQL over textual and database data together. We show how real pharmacological queries can be answered over 4 billion text mined triples.

Keywords: text mining, NER, retrieval, RDF, SPARQL, Open PHACTS

1 Introduction

The process of developing and discovering new drugs (i.e. pharmacology) requires deep analysis of the relationships between chemical and biological information. Currently, scientists use a multitude of information sources ranging from databases (e.g. Drug Bank[1]) to specialized literature search services (e.g. PubMed[2]) to standard web search. A typical pharmacological analysis may require the consultation of up to 8 data sources. This poses a tremendous burden on the researcher who has to deal with a wide range of incompatible and often inconsistent data sources. To address this problem, the Open PHACTS (Open Pharmacological Concepts Triple Store[3]) project aims to deliver a single view across available pharmacological information sources. Several approaches integrating life science data using Semantic Web technologies have been described in the literature [1][2][3]. However, these approaches have largely ignored the vast amount of content only available within the scientific literature. Results from text are vital as they often offer more up-to-date information than found in curated databases as well as have information that might not be included in such

1 http://www.drugbank.ca/
2 http://www.ncbi.nlm.nih.gov/pubmed/
3 http://www.openphacts.org/

databases. A key goal of Open PHACTS is to provide the capability to query over textual and database information together.

Because of the complex relational nature of pharmacological queries, this goal cannot be achieved through standard information retrieval mechanisms (i.e. keyword or Boolean search). Instead, textual information needs to be presented in a *structured* fashion that can be integrated with other structure data sources (e.g. databases). In this work, we present an RDF schema for text mining results that enables such complex relational queries represented in SPARQL to be performed. We show how real pharmacological queries can be answered over 4 billion text mined triples. We now describe the developed schema for text mining results. We then discuss the usage of that schema to answer an exemplar pharmacological question.

2 The Text Mining Schema

As a starting point the complete PubMed containing more than 19 million abstracts has been processed with the text mining tool ProMiner [4]. ProMiner is a dictionary based NER approach which allows extracting a large variety of semantic entities, i.e. genes/proteins, drugs, chemical names, organisms and diseases. It can be combined with pattern matching and machine learning approaches extracting chromosomal locations, rs numbers, SNPs, IUPAC expressions and epigenetic modifications. The following data sources have been used:

- PRT files: Contain information about the matched entities (hits) and the found synonyms, positional information as well as scoring information for machine learning based filtering.
- MAP files: Contain references to external data sources for each concept as well as a preferred label which is used for visualisation.
- SYN files: Dictionaries used by ProMiner to find all matches in text.
- Sentence files: Sentence based full text files for documents in PubMed.

In the following section, the RDF schema for text mining results (TMS), shown in figure 1, is described. TMS consists of three main parts: text mining data, document database and mapping/concept store. In principle all data can be represented in a single connected graph. We have decided to represent them in different graphs in order to add provenance data to them. We can for example distinguish between different runs of the text mining tools and between different versions of dictionaries.

The starting point of the schema is the *hit*. I can be found in the center figure 1. A hit describes a text mining match of a concept at a specific position in a document. Thereby the hit node is connected with the document and the concept node, as well as a position node which stores the offset in the document and its definition. The hit node is connected via *run* to the annotator node which represents the generator of the hit. In our case this source is ProMiner in version 3.7. The hit node is associated with a confidence, which is necessary to describe the quality of the automatic generated hit. The *document node* is another main node of the schema. It represents the document in which a hit is found and is connected to several nodes representing its title, abstract, etc. In our case we are using *pmids* as a URI for a document. The document

node is also connected to the sentence node. The sentence node contains the original sentences from the document, allowing full text search making use of the regular expression filter of SPARQL.

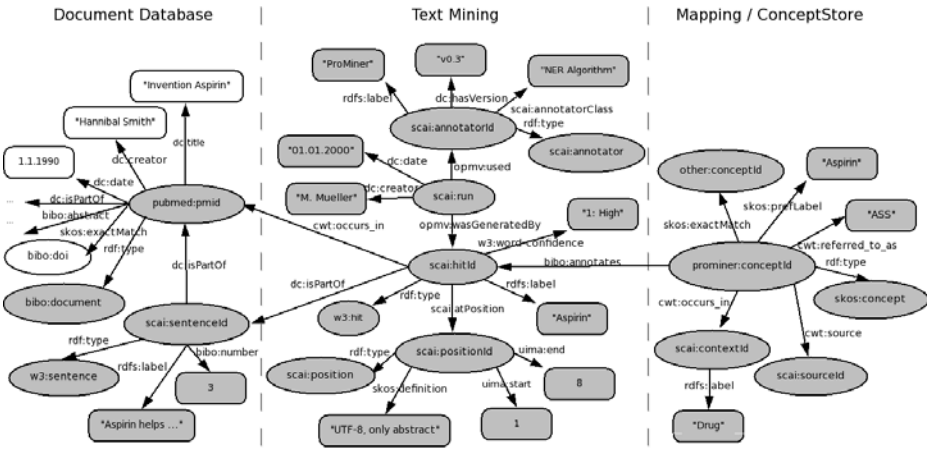

Fig. 1. The TMS schema for text mining results is shown as a labeled graph. The nodes describe subjects and objects in RDF. The round ones thereby represent URIs and the squared ones literals. The edges between the nodes represent the predicates which describe the relation between two nodes. Grey nodes have been populated with ProMiner data. White nodes contain bibliometric information from PubMed.

The last main node is the concept node. A concept can be a specific protein for example. The concept node just describes it in an abstract way, due to the absence of a unique naming convention. Concepts between different data sources are mapped through the skos:*exactMatch* relationship. All possible names are stored in the synonym node. Therefore any specific name from the synonyms can be used in a SPARQL query to include the other synonyms by using the concept. Next to the synonyms also the context of a concept is stored. In our example the context would be 'Protein'. Concept hierarchies can be modeled by making use of the rdfs:*subclassOf* predicate. This allows us to include all subclasses of a concept in a query.

Next to the specific nodes mentioned above there exist two node types which are always attached by the standard predicates rdfs:*label* and rdf:*type*. The label nodes provide a human readable version of the URI of the connected node, and should mainly be used for visualization. The type nodes are used to address a specific node of the schema in a query. Without these nodes it would not be possible to distinguish between e.g. a hit and a document node in a query.

We designed all relations as unidirectional, but in principle they can also be bidirectional. In our case most relations are 1:n. The relation between a hit and a concept can be n:m. A concept can have several hits, but also to one hit several concepts can be assigned. This happens if there are ambiguities which cannot be resolved by text mining (e.g. gene or protein).

3 Use Case Query and Conclusion

The Open PHACTS project in consultation with biologists and chemists developed a prioritized list of 83 research questions. These questions revolve around pharmacological concepts namely genes, pathways, diseases, targets, and drugs. Typical questions are “For my specific target, which active compounds have been reported in the literature?”, “For my given compound, which targets have been patented in the context of Alzheimer's disease?” We use at the highest priority question (Q15) as an exemplar:

“Find me all oxidoreductase inhibitors which are active in both human and mouse (IC50 < 100nM).”

Note that it took a small team of experts at USC[4] two days to answer this query by manually aggregating data about compound ID, chemical structure, target, species, mode of action, IC50, and assays. We formulated Q15 as a SPARQL query and executed it over our database. The answer set contains 561 unique triples for the database part and 60 triples referring to publications related to this query.

Our approach shows that structured representation of text mining results can enable complex relational queries to be answered for the pharmacological domain. We see TMS as a core artifact in the goal to enable complex queries over the combination of textual and structured data sources.

Acknowledgments. We would like to thank the USC team for providing us with example data for Q15 and the SCAI team for providing us with the text mining data. We acknowledge financial support provided by the IMI-JU, grant agreement n° 115191.

References

1. Chen, B., Dong. X., Jiao, D., Wang, H., Zhu, Q., Ding, Y., Wild, D.J.: Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data. BMC Bioinformatics 2010, 11, pp. 255-
2. Samwald, M., Jentzsch, A., Bouton, C., Kallesøe, C.S., Willighagen, E., Hajagos, J., Marshall, M.S., Prud'hommeaux, E., Hassanzadeh, O., Pichler, E., Stephens, S.: Linked open drug data for pharmaceutical research and development. Journal of Cheminformatics 2011, 3, pp. 19-
3. Willighagen, E.L., Alvarsson, J., Andersson, A., Eklund, M., Lampa, S., Lapins, M., Spjuth, O., Wikberg, J.E.S.: Linking the Resource Description Framework to cheminformatics and proteochemometrics. Journal of Biomedical Semantics 2011, 2(Suppl 1), pp. 6-
4. Hanisch, D., Fundel, K., Mevissen, H.T., Zimmer, R., Fluck, J.: ProMiner: Rule based protein and gene entity recognition. BMC Bioinformatics 2005, 6(Suppl 1):pp. 14-

4 Grupo BioFarma-USEF. Departamento de Farmacología. Facultad de Farmacia. Campus Universitario Sur s/n Santiago de Compostela. http://www.usc.es/biofarma/