Semantic Web technologies for a knowledge base of biomedical facts extracted from scientific literature

Maria Biryukov[1], Valentin Groues[1], Christophe Trefois[1], Venkata Satagopam[1], and Reinhard Schneider[1]

Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, Esch-Belval, Luxembourg

1 Objective

Biomedical literature, including scientific articles, public health reports and books become more and more available due to massive digitalization. Exploration and analysis of this rich source of data requires assistance of automatic tools capable of dealing with large volumes of text. We are developing a pipeline for processing publicly available biomedical text, abstracts, full text articles, conference proceedings, eventually books and electronic health records, starting from searching the web and downloading raw files, to extraction and storing concepts (entities) and semantic relations between them into a knowledge base. The goal is to create a biomedical knowledge base publicly available for both human and machine access (SPARQL endpoint and REST API).

2 Events Extraction Pipeline

For the extraction of biomedical entities from the literature, we rely first on Reflect (httpd://reflect.ws) - a named entity recognition engine to identify biomedical concepts in the text. GeniaTagger is used to obtain basic morphological and syntactic information. The latter is completed by application of the Stanford Syntactic Parser which converts sentences into syntactic trees representing dependencies between the words. Combined with a set of rules and dedicated patterns, this information allows for getting semantic interpretation of sentences and extraction of meaningful relationships between the concepts.

3 Semantic Web technologies

An ontology has been created to represent the biomedical events extracted from the literature. Examples of extracted biomedical events include, among others, proteins interactions, chemicals effects and genes expression. Named graphs are used to add metadata about those events (e.g. scoring). The events are stored in a triple store (OpenLink Virtuoso). A hierarchy of biomedical relationships has been defined and the reasoning capabilities of Virtuoso are used to dynamically

use this hierarchy at query time. From 1 million publications already processed, 11 millions events were extracted resulting in about 150 millions of triples. A first prototype of a web-based visualization tool has been created to browse this knowledge base.