WP7 Semantic Infrastructure & Prototype Building

by Ontotext AD

Overall Demo Scenario

The Demo for the patent use case of WP7 will be based on Exopatent (http:\\exopatent.ontotext.com) and already built KRI infrastructure (http:\\molto.ontotext.com).

It will include the following search options incorporated into a single interface:

- Natural language search of structured data (in similar way as in KRI demo).
- Patterns, which will appear in the language of the interface with the possibility to switch the language. (Optional)
- Facets, which will allow queries in the languages of MOLTO project (Optional)
- Boolean, which will allow queries in the languages of MOLTO project (Optional)
- Mimir search, which will allow to query in the languages of MOLTO project (Optional)

This will allow querying about claims, drugs that are subject of the patent, times when the patent was filed, etc.

The language of the interface and queries will be automatically chosen by the locale of the browser but the user will be able to change the language in which he would like to make a query.

Architecture

We will maintain several indexes, e.g. for the metadata and for the document contents. English will be the basic processing language.

Each incoming query will be translated into English, search in the document indexes will be performed if the search option requires search in the document content indices. The language of the query will be passed on too. the translated documents will be stored in a folder per language, and connected with the RDF indexes. The documents in the language of the query will be selected and returned to the user.

Labels of the triples will also be provided in the languages of MOLTO, and will be retrieved according to the language of the query.

The main technologies that are planned to be used are Owlim(http://www.ontotext.com/owlim), Gate(http://gate.ac.uk/), Mimir(http://gate.ac.uk/family/mimir.html), Lucene(http://lucene.apache.org), Kim(http://www.ontotext.com/kim).

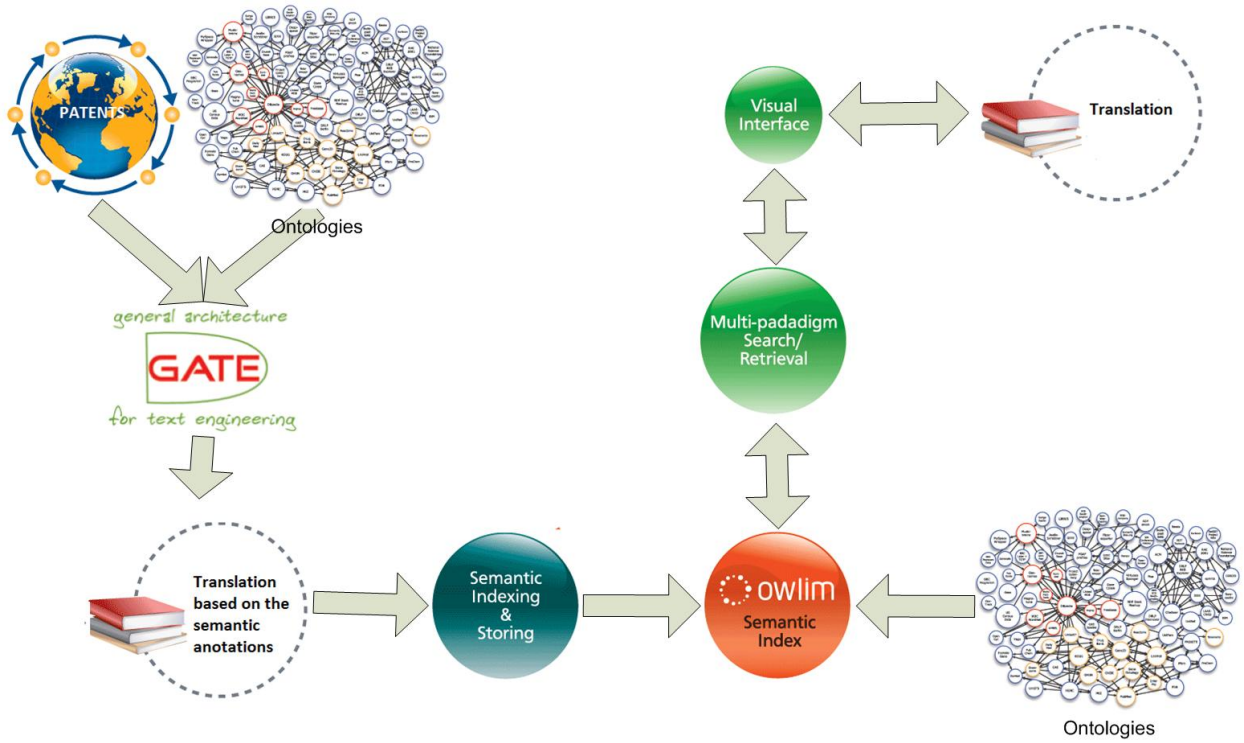

Figure 1. Architecture of the system

Data Models

Ontotext will use its experience in semantic annotation to annotate the patents provided by the partners. The main ontologies are showed on figure 2 and figure 3.

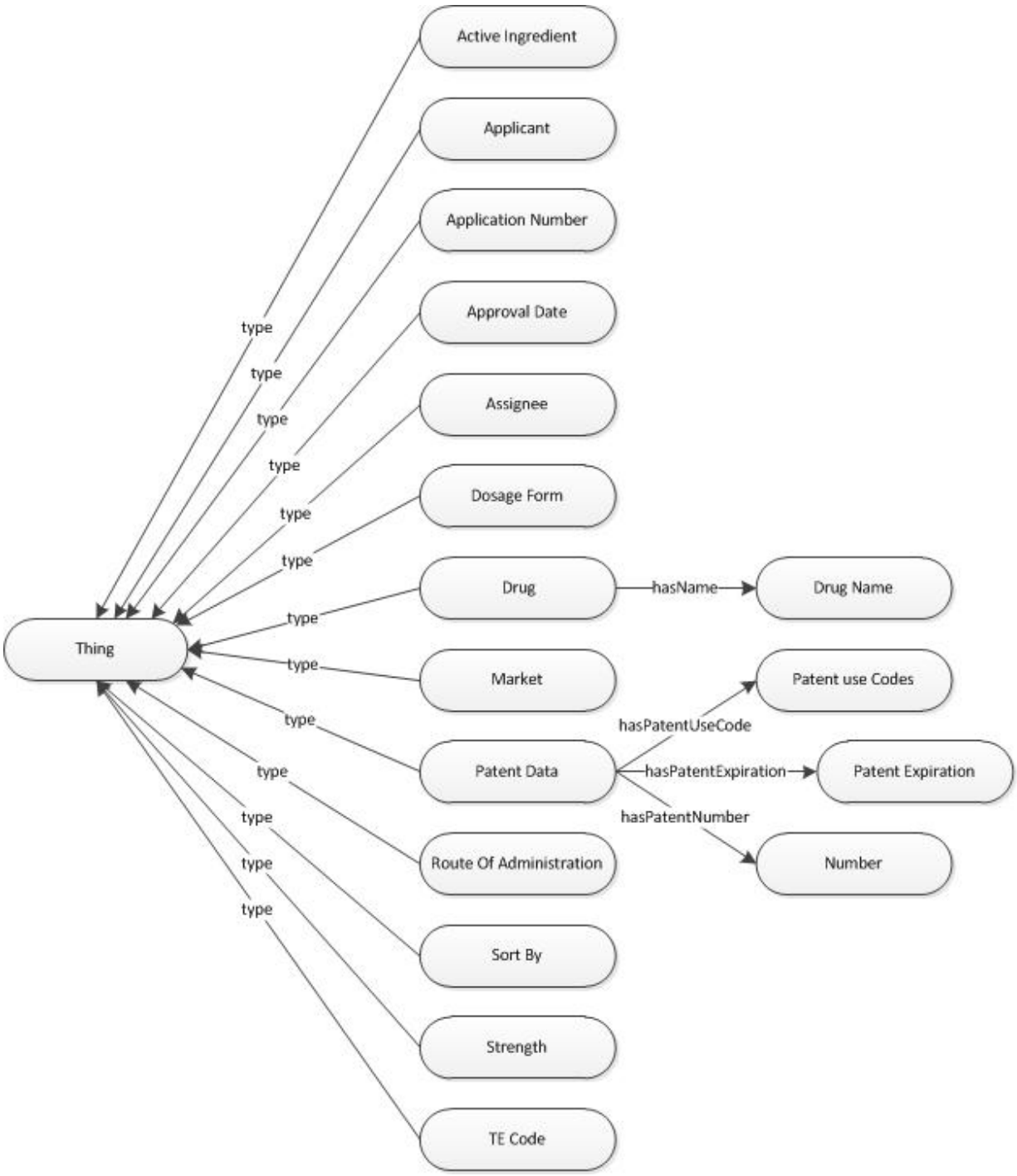

Figure 2. Ontology 1

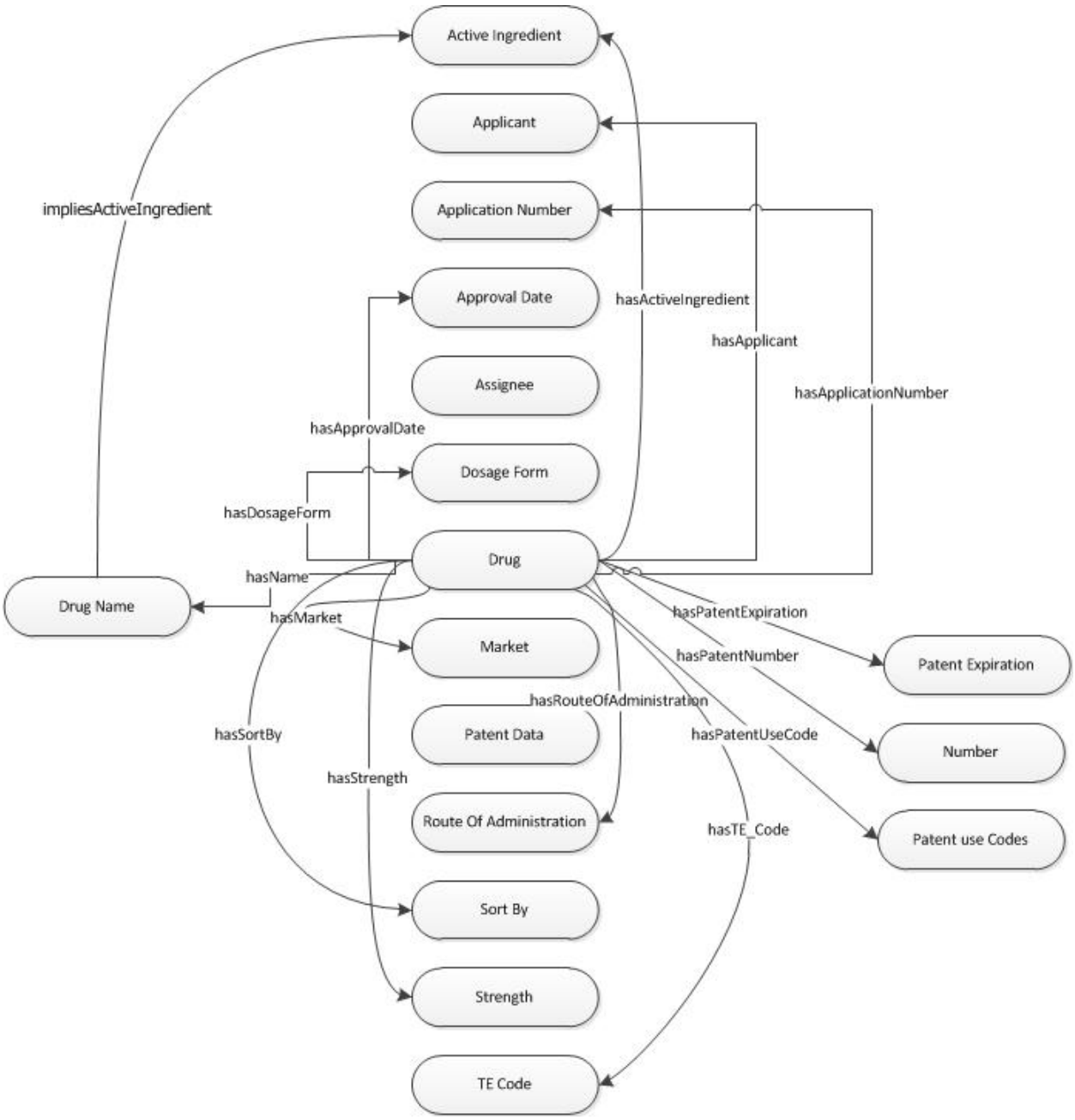

Figure 3.Ontology 2

The annotated documents by the above ontologies are showed on Figure 4.

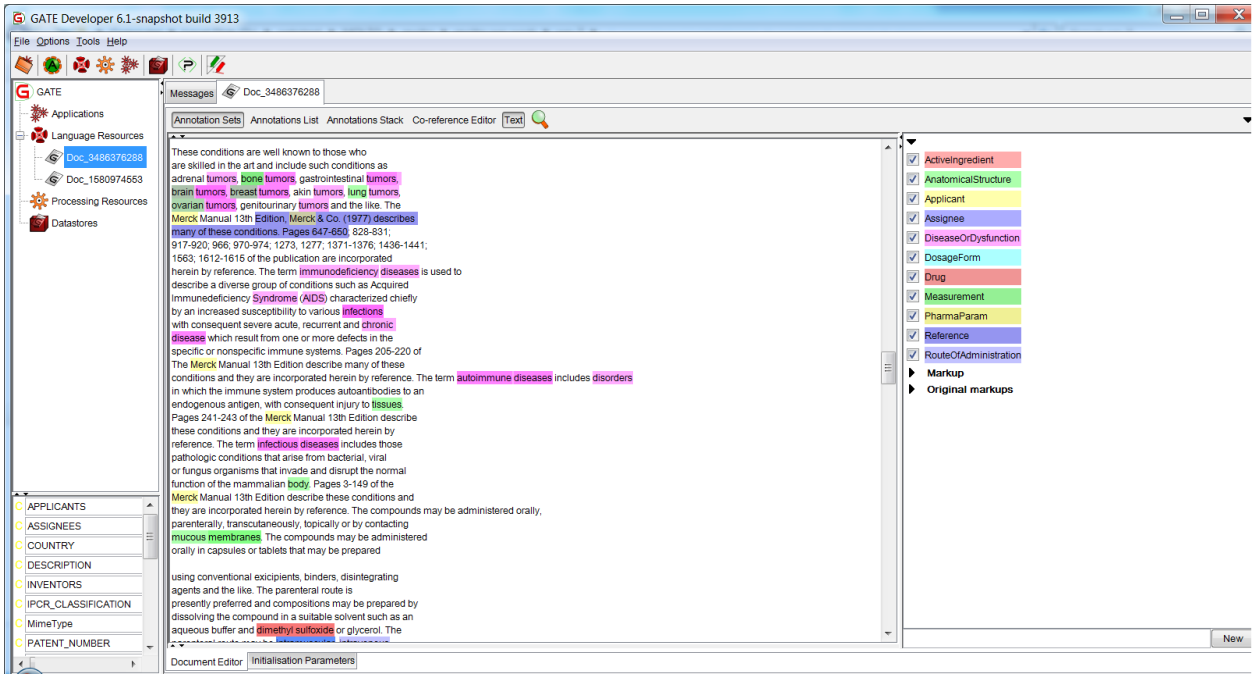

Figure 4. Anotated document

Natural Language Queries - TODO! All possible queries to be defined as restricted language.(UPC and Ontotext) Make correspondent GF grammar(UGOT) and generate GF Trees that to be semi-automated translated to SPARQL(Ontotext).

Facet search, Boolean search, Keyword search, Mimir search - TODO! Translation of the keywords in the different languages.

WP7 Project phases until September 2011

(1) models: ontology and queries
(2) backend: indexes composition: mimir index, kim index, owlim triples

Deadline: July 31st

(3) translation facilities – queries, documents, labels
(4) data integration

Deadline: August 15th

(5) user interfaces in MOLTO languages
(6) integration

Deadline: September 30th