D7.2 Patent MT and Retrieval Prototype

Contract No.:	FP7-ICT-247914
Project full title:	MOLTO - Multilingual Online Translation
Deliverable:	D7.2 Patent MT and Retrieval Prototype
Security (distribution level):	Public, regular publication
Contractual date of delivery:	M27
Actual date of delivery:	September 2012
Type:	Prototype
Status & version:	final 1.2
Author(s):	Meritxell Gonzàlez[1], Milen Chechev[2], Mariana Damova[2], Ramona Enache[3], Cristina España-Bonet[1], Lluís Màrquez[1], Maria Mateva[2], Aarne Ranta[3], Laura Tolo̧si[2]
Task responsible:	UPC[1]
Other contributors:	Ontotext[2], UGOT[3]

ABSTRACT

The present document is Deliverable D7.2 of WP7. It gives a description of the multilingual patents retrieval prototype produced in this workpackage and a brief user manual to access the demo.

The main highlights achieved in the prototype with respect to the beta version described in the Deliverable 7.1[CEEB+12] are the following: a) The demo allows for querying the system in the three languages addressed in this WP (English, French and German); b) the patents in the database has original text in English, French and German and also the translated documents for all missing languages of each document; c) the patent document translation can be done following a simple pipeline; d) some improvements on the interface addressed several deficiencies detected during internal evaluation; e) the new query library and its application to the patents use case have been presented at the Third Workshop on Controlled Natural Language (CNL 2012[1]), being held in Zurich at the end of August 2012.

[1]http://attempto.ifi.uzh.ch/site/cnl2012/

Contents

1 Introduction

This document corresponds to the second Deliverable of WP7: ``Patents Case Study''. It describes the multilingual patents retrieval prototype and the technologies and resources that it integrates. The last section contains also a brief user manual to access the online interface, which is publicly available at:

http://molto-patents.ontotext.com/.

The purpose of WP7 is to tackle a MOLTO case study centered on the patents domain. This case study aims to create a prototype for automatic translation and retrieval of patents, allowing robust translation of patent abstracts and claims, cross-language retrieval of patent data and multilingual queries.

The prototype is publicly available and it can be accessed at the mentioned URL. The preliminary version of the prototype, described in Deliverable 7.1 [CEEB+12] had only original patent documents in the databases and the system was only available in English and French. The present version of the prototype allows for querying also in German. Moreover, the controlled natural language covered by the query grammars has been revised using the new Query Library and the tools developed in WP4. With respect to the documents, we have integrated a larger dataset of patents (see Section 2.1). It has been completely translated using an Statistical Machine Translation (SMT) system trained on the domain. Nonetheless, by the time of the final report, we will translate them using the hybrid system that is being developed within WP5.

The recommendations given in the 2nd year review have been also addressed or are part of our work in progress. With respect to semantic annotation, it was unclear how the use of different resources (i.e., overlaps may need for coordination) was addressed. This issue is discussed in Section 2.3. The evaluation of the different modules and technologies involved in the prototype have been included in D9.1.E. The goal of transferring semantic annotations to the target language is our current work in progress in which we are updating the pipeline discussed in Section 2.1. In relation to the grammar – ontology interoperability automation, it has been addressed as part of WP4 work, and a specific evaluation for applied to this WP7 is part of our work for the final report.

2 Prototype overview

This section gives a general description of the patents prototype. It is centered on the resources that are used and generated by the modules of the system and how they are integrated in the system. The resources described in this document are all available at the MOLTO repository[2].

The multilingual patents retrieval prototype consists mainly of four modules (see Figure 1). The patent documents are preprocessed and translated using a statistical system trained on the biomedical domain (see Section 2.1). The original and translated documents are used to feed the retrieval system following the process described in Section 2.2

[2]The MOLTO repository is hosted at UGOT facilities. Access is granted for all MOLTO members.

and Section 2.3. Users can access the system through an online interface that allows for querying the system using a controlled natural language (CNL).The queries are processed using a GF query grammar that have been adapted for the patents domain. This grammar, described in Section 2.4, follows the general query library developed in WP4.

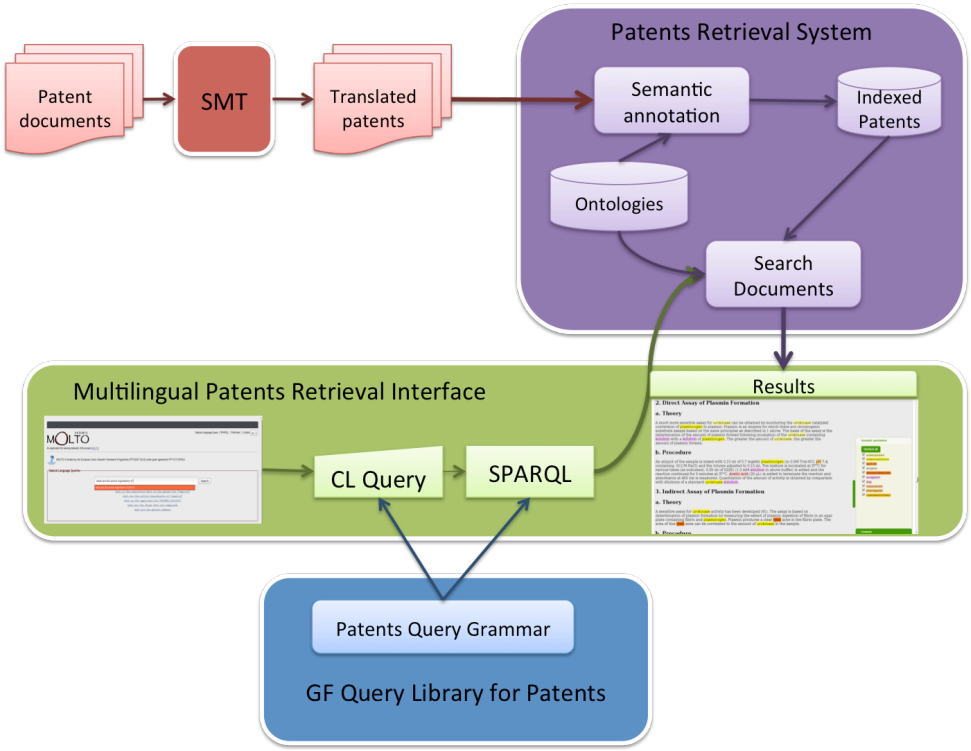

Figure 1: General architecture of the prototype

2.1 Patent corpus and Translation

The preparation of the patent corpus and the translation of the documents is part of the work carried at UPC. For the patents case study we obtained two different datasets. On the one hand, the European Patent Office (EPO[3]) provided some parallel corpus containing the text of 66 patents belonging to the biomedical domain (IPC A61P). This corpus, which only contains the parallel raw text and the identifier of the patent, is being used as the test set of the translation systems developed in WP5.

On the other hand, EPO provided also a website from where we downloaded 7,705 patent documents, also in the biomedical domain, all dated from 2010 to 2012. The patent documents follow the normalized XML format defined by the EPO. In general, this format consists of the following sections: bibliographic data, abstract, description, claims, and references. The abstract, the description and the claims are always written in one of the

[3] http://www.epo.org/

three official languages, i.e., English (EN), German (DE) and French (FR), and sometimes they contain also the translation to any of the other two languages or both of them. In our dataset, up to 4,274 out of the 7,705 documents have claims, and 2,058 out of them are trilingual. 2,116 documents have claims written only in English, 66 have claims only in German, 34 only in French. Table 1 gives a general overview of the number of sections in the corpus and the languages in which they are written.

	English	German	French
Claims	4,174	2,124	2,092
Abstracts	2,552	83	45
Descriptions	3,937	201	136

Table 1: Number of sections and languages in the corpus of patents

Due the characteristics of these documents, they do not constitute an aligned corpus and, in consequence, they cannot be used for training the SMT systems (which are trained using the dataset described in Deliverable 5.1 [EBGM11]). Instead, we are using these documents to feed the patents retrieval system. To this end, the patents are automatically translated using the process described below and semantically annotated using the process described in Section 2.3. The complete collection of files is available in the MOLTO repository[4], and it consists of 1) the original patent documents, 2) the English version of the patent documents having the semantic annotations, and 3) the automatic translations of claims, abstracts and descriptions. Table 2 gives a numerical description of the dataset, i.e., the number of documents, segments and tokens in English, German and French.

	Documents	Segments	Tokens
English	6,431	9,582,864	178,213,580
German	2,276	306,495	4,811,281
French	2,205	210,739	3,892,813

Table 2: Numerical description of the patents dataset

2.1.1 Translation of the documents

The designed process for patents translation allows for building a translated document having the same XML structure as the original patent. As a result, the interface of the prototype can show the translated patents using the same user-friendly view as for the original ones.

The pipeline of the process is shown in Figure 2 and the example below (see Section 2.1.2) shows the transformations on the text at each step. The first step shows the

[4] svn://molto-project.eu/patents-corpora/EPO-www-patents/

original content of a patent document. The excerpt in the examples belongs to the 17th paragraph of the English description of the patent number EP1330442B1. It contains several especial sections such as image, listings, subindexes and comments. As shown in the diagram, the patent files are preprocessed in order to extract the text contained into the sections in a structured manner. First, the formatting marks inline with the text are replaced by placeholders (step 2). And then, the resulting text is segmented and tokenized as required by the translation system (step 3). After this step the structural marks have been removed and the remaining consists of raw text having the placeholders. Soon after, the raw text is translated using the SMT system (step 4). The translated text is post-processed in order to recover the original structure of the document (step 5), including original formatting, claims enumeration and images. To this end, the process uses the original XML document.

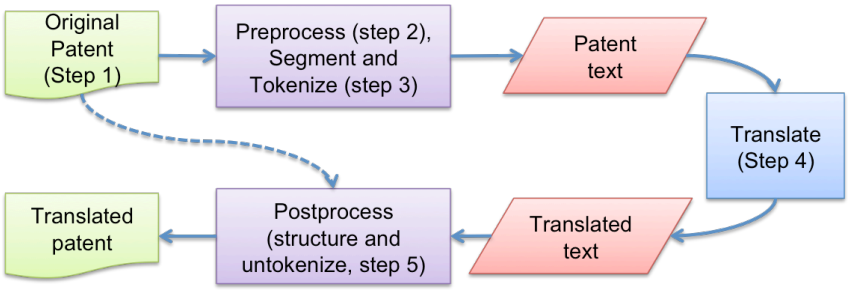

Figure 2: Patent document translation pipeline

The patent documents are translated using the SMT system described in Deliverable 5.2 [MOL12]. The current version of the prototype uses a phrase-based system adapted to and trained on parallel patents in the biomedical domain (see Deliverable 5.1 [EBGM11] for more references about the corpus). The SMT system has been built using standard freely available software. A 5-gram language model is estimated using interpolated Kneser-Ney discounting with SRILM [Sto02]. Word alignment is done with GIZA++ [ON03] and both phrase extraction and decoding are done with the Moses package [KSF+06, KHM+07]. The optimization of the weights of the model is trained with MERT [Och03] against the BLEU [PRWZ02] evaluation metric.

The source code for the rest of the pipeline is available at the MOLTO repository[5]. In order to facilitate its use, two main scripts perform all the needed calls sequentially. One of them is used to process and translate a single file, while the other one can translate a bunch of files, all from the same source-target pair of languages, and it is optimized to parallelize the processes if an appropriate computational environment is available. Further instructions about how to use the scripts are given in a README file along with the source code.

[5]The source files can be found in `svn://molto-project.eu/patents-corpora/corpora-parser.tgz`.

2.1.2 Example showing the transformation steps needed to translate a excerpt of text.

1. The original text extracted from the patent number EP1330442B1.

```
<p id="p0017" num="0017">A third aspect of the present invention relates to a
    pharmaceutical composition comprising a compound of the formula:
<chemistry id="chem0003" num="0003"><img id="ib0003" file="imgb0003.tif" wi="53" he="41"
    img content="chem" img format="tif"/></chemistry>
or isomers, salts, solvates and chemically protected forms thereof, wherein:
<ul id="ul0002" list style="none" compact="compact">
<li>A and B together represent a fused aromatic ring, optionally substituted with one or
    more substituent groups selected from halo, nitro, hydroxy, ether, thiol, thioether,
    amino, C<sub>1 7</sub> alkyl, C<sub>3 20</sub> heterocyclyl and C<sub>5 20</sub> aryl
    ;</li>
<li>R<sub>C</sub> is  CH<sub>2</sub> R<sub>L</sub>;</li>
<li>R<sub>L</sub> is phenyl optionally substituted with one or more substituent groups
    selected from C<sub>1 7</sub> alkyl, C<sub>5 20</sub> aryl, C<sub>3 20</sub>
    heterocyclyl, halo, hydroxy, ether, nitro, cyano, carboxy, ester, amido, amino,
    sulfonamido, acylamido, ureido, acyloxy, thiol, thioether, sulfoxide and sulfone; and
     R<sub>N</sub> is hydrogen,</li>
<li>and a pharmaceutically acceptable carrier or diluent.</li>
</ul><!  EPO <DP n="6">  ></p>
```

2. The pre-processed text.

```
<p id="p0017" num="0017">A third aspect of the present invention relates to a
    pharmaceutical composition comprising a compound of the formula:
<chemistry id="chem0003" num="0003"><img id="ib0003" file="imgb0003.tif" wi="53" he="41"
    img content="chem" img format="tif"/></chemistry>
or isomers, salts, solvates and chemically protected forms thereof, wherein:
<ul id="ul0002" list style="none" compact="compact">
__LI__A and B together represent a fused aromatic ring, optionally substituted with one
    or more substituent groups selected from halo, nitro, hydroxy, ether, thiol,
    thioether, amino, C__SUB__1 7__/SUB__ alkyl, C__SUB__3 20__/SUB__ heterocyclyl and
    C__SUB__5 20__/SUB__ aryl;__/LI__
__LI__R__SUB__C__/SUB__ is  CH__SUB__2__/SUB__ R__SUB__L__/SUB__;__/LI__
__LI__R__SUB__L__/SUB__ is phenyl optionally substituted with one or more substituent
    groups selected from C__SUB__1 7__/SUB__ alkyl, C__SUB__5 20__/SUB__ aryl, C__SUB__3
     20__/SUB__ heterocyclyl, halo, hydroxy, ether, nitro, cyano, carboxy, ester, amido,
    amino, sulfonamido, acylamido, ureido, acyloxy, thiol, thioether, sulfoxide and
    sulfone; and R__SUB__N__/SUB__ is hydrogen,__/LI__
__LI__and a pharmaceutically acceptable carrier or diluent.__/LI__
</ul><!  EPO <DP n="6">  ></p>
```

3. The raw after segmentation and tokenization.

```
A third aspect of the present invention relates to a pharmaceutical composition
    comprising a compound of the formula :
or isomers , salts , solvates and chemically protected forms thereof , wherein :
__LI__ A and B together represent a fused aromatic ring , optionally substituted with one
     or more substituent groups selected from halo , nitro , hydroxy , ether , thiol ,
    thioether , amino , C__SUB__1 7__/SUB__ alkyl , C__SUB__3 20__/SUB__ heterocyclyl and
     C__SUB__5 20__/SUB__ aryl ; __/LI__ __LI__R__SUB__C__/SUB__ is  CH__SUB__2__/SUB__
    R__SUB__L__/SUB__ ; __/LI__ __LI__R__SUB__L__/SUB__ is phenyl optionally substituted
    with one or more substituent groups selected from C__SUB__1 7__/SUB__ alkyl ,
    C__SUB__5 20__/SUB__ aryl , C__SUB__3 20__/SUB__ heterocyclyl , halo , hydroxy ,
```

```
    ether , nitro , cyano , carboxy , ester , amido , amino , sulfonamido , acylamido ,
    ureido , acyloxy , thiol , thioether , sulfoxide and sulfone ; and R__SUB__N__/SUB__
    is hydrogen , __/LI__ __LI__and a pharmaceutically acceptable carrier or diluent . __
    /LI__
```

4. The text translated into French using the SMT system.

```
Un troisi aspect selon la prnte invention se rapporte ne composition pharmaceutique
    comprenant un compose la formule :
ou isoms , sels , solvates et chimiquement ses formes protes , dans laquelle :
__LI__A et B forment ensemble un noyau aromatique fondu , ntuellement substituar un ou
    plusieurs groupes substituants choisis parmi les substituants halo , nitro , hydroxy
    , thiol , er d'alkyle en C1 10 , thioer , amino , alkyle C__SUB__1 7__/SUB__ ,
    C__SUB__3 20__/SUB__ hrocyclyle et C__SUB__5 20__/SUB__ aryl ; __/LI__
    __LI__R__SUB__C__/SUB__ est CH__SUB__2__/SUB__ R__SUB__L__/SUB__ ; __/LI__
    __LI__R__SUB__L__/SUB__ est un phle ntuellement substituar un ou plusieurs groupes
    substituants choisis parmi C__SUB__1 7__/SUB__ alkyle , aryle , hrocyclyle C__SUB__5
     20__/SUB__ C__SUB__3 20__/SUB__ , halo , hydroxy , er d'alkyle en C1 10 , nitro ,
    cyano , carboxy , ester , amido , amino , sulfonamido , acylamido , uro , acyloxy ,
    thiol , thioer , dimylsulfoxyde et sulfone ; et R__SUB__N__/SUB__ est hydrogen , __/
    LI__ __LI__and un support acceptable du point de vue pharmaceutique ou diluent . __/
    LI__
```

5. The excerpt fit into the original XML structure.

```
<p num="0017" id="p0017" >
Un troisi aspect selon la prnte invention se rapporte ne composition pharmaceutique
    comprenant un compose la formule :
<chemistry num="0003" id="chem0003" >
<img file="imgb0003.tif" he="41" id="ib0003" img content="chem" img format="tif" wi="53"/
    >
</chemistry>
<ul compact="compact" list style="none" id="ul0002" >
ou isoms , sels , solvates et chimiquement ses formes protes , dans laquelle :
<li>A et B forment ensemble un noyau aromatique fondu , ntuellement substituar un ou
    plusieurs groupes substituants choisis parmi les substituants halo , nitro , hydroxy
    , thiol , er d'alkyle en C1 10 , thioer , amino , alkyle C<sub>1 7</sub> , C<sub
    >3 20</sub> hrocyclyle et C<sub>5 20</sub> aryl;</li> <li>R<sub>C</sub> est CH<sub
    >2</sub> R<sub>L</sub>;</li> <li>R<sub>L</sub> est un phle ntuellement substituar un
     ou plusieurs groupes substituants choisis parmi C<sub>1 7</sub> alkyle , aryle ,
     hrocyclyle C<sub>5 20</sub> C<sub>3 20</sub> , halo , hydroxy , er d'alkyle en C1 10
     , nitro , cyano , carboxy , ester , amido , amino , sulfonamido , acylamido , uro ,
    acyloxy , thiol , thioer , dimylsulfoxyde et sulfone ; et R<sub>N</sub> est hydrogen
    ,</li> <li>and un support acceptable du point de vue pharmaceutique ou diluent.</li>
</ul>
</p>
```

2.2 Patents Retrieval system

The patent retrieval prototype is an adaptation to the patent domain of the retrieval system developed in WP4. This system, developed and adapted by Ontotext, combines machine translation and retrieval of patents in the biomedical and pharmaceutical domains. It provides an interface for natural language queries in 3 languages (English, German and French) and the potential to retrieve results from both structural knowledge databases

(ontologies) and multilingual documents (patents). As mentioned, the patent retrieval prototype uses the infrastructure that is defined in the Deliverable 4.1 [MI10] and its functionality has been extended by adding document indexing and retrieval.

The web interface of the patent retrieval prototype is made as an overlay of the WP4 prototype as it has been specialized for the patents use case and the patent documents described in Section 2.1. To this end, specific actions and new functionalities were added to the patent prototype, such as document indexing and semantic annotation, biomedical ontologies (see Section 2.3), patent query language (see Section 2.4 and document visualization (see Section 3).

Document indexing.

The WP4 prototype uses only the semantic data loaded at the OWLIM semantic repository[6]. The WP7 is focused on the patents domain and, in consequence, the prototype provides the ability to search patents and to retrieve complete documents.

Document annotation.

The documents are semantically annotated in order to attach the semantic concepts to the terms that are contained in the text of the patent. For matching purposes, the semantic annotations are linked to the document identifier (e.g., the patent number) and stored at the semantic repository. The excerpt of text in Figure 3 shows an example of an annotated paragraph. The tag *DiseaseOrDisfunction* is used to add the information about the semantic instance and its class. Once we have documents that are annotated and their content is connected with the semantic classes, the system is able to search for patents that contain a specific concept, such as a drug, disease or active ingredient.

```
</p>It is clear that the compound having the above-mentioned activity ameliorates the
<DiseaseOrDysfunction gate:gateId="45427" inst="http://linkedlifedata.com/resource/umls/id/C0233794"
 class="http://linkedlifedata.com/resource/semanticnetwork/id/T046">memory deficits
</DiseaseOrDysfunction> (i.e. <DiseaseOrDysfunction gate:gateId="45430" inst=
"http://linkedlifedata.com/resource/umls/id/C0002622" class=
"http://linkedlifedata.com/resource/semanticnetwork/id/T046">amnesia</DiseaseOrDysfunction>,
<DiseaseOrDysfunction gate:gateId="45429" inst="http://linkedlifedata.com/resource/umls/id/C0497327"
 class="http://linkedlifedata.com/resource/semanticnetwork/id/T046">dementia</DiseaseOrDysfunction>
, etc.) from the description in the Journal of Pharmacology and Experimental Therapeutics, Vo. 279,
No. 3, 1157-1173 (1996). Further, it is expected that the compound having the above-mentioned
activity is useful as therapeutical and/or preventive agent for aforesaid <DiseaseOrDysfunction
gate:gateId="45411" inst="http://linkedlifedata.com/resource/umls/id/C0012634" class=
"http://linkedlifedata.com/resource/semanticnetwork/id/T046">diseases</DiseaseOrDysfunction> from
some patent applications (e.g. PCT International Publication No. WO 98/27930, etc.).</p>
```

Figure 3: An excerpt of text having semantic annotations

Document visualization.

The online interface allows the user to access the retrieval system, execute queries and obtain the results in a browsable fashion. Furthermore, the user can select any of

[6] http://www.ontotext.com/owlim

the available languages and browse the results according to the selection. The results obtained consists of the ontologies' instances of the query that are matched in the semantic repository and the set of documents that are related to these instances. Both, the collection of instances and the documents can be navigated from the user interface.

Visualization of the annotations at the document.

For convenience of the user, the semantic annotations are highlighted on the document. The different types of annotations are marked with several colors in order to improve the readability and friendliness of the document. An additional functionality in the interface allows the user to select just concrete classes of annotation and hyperlinks from the semantic annotations to the semantic instances in the repository.

Specific query language.

The query language defined for the prototype developed in WP4 covers the upper level domain described in the PROTON ontology[7]. Its concepts describe people, locations, institutions, the most popular named entities that are usually looked for. For the patent use case we needed a more specific query language so it has been adapted to cover questions in the biomedical domain.

Biomedical ontologies added to the database.

Because of the topic of the use case the ontologies that are loaded to the prototype differ from the ontologies in WP4. They describe concepts of the biomedical domain and the patents structure. The next Section 2.3 gives a more detailed description of them. Besides, the annotation process is made using the GATE framework and the customized pipeline for patent annotation. The annotations are produced based on gazetteers populated from the ontology resources and then, then they are used to search for and retrieval of the patent documents.

2.3 Ontologies and Document Indexing

The main goal of the semantic retrieval system is to enable users to obtain information about concepts, alias entities, that are found in documents. To achieve this, it is necessary to have a structured semantic representation of the concepts. This structured semantic representation is called ontology. Ontologies represent strictly defined concepts and the relationships between them. They allow new knowledge to be derived based on their representations and the explicit facts available in the knowledge base. For instance, one can have explicit information that *ampicillin is an FDA Drug*, and that an *FDA Drug is a Drug*, so additional information can be generated saying that *ampicillin is a Drug*. On the one hand, ontologies are used during the process of semantic annotation in order to link the language expressions with semantically identifiable units. On the other hand, they are also used to provide connection with the biomedical semantic knowledge bases (cf., the

[7] http://proton.semanticweb.org/

Ontotext service http://linkedlifedata.com) that provide extensive information about the concepts.

The prototype described in this deliverable implements the knowledge representation infrastructure built in WP4, and described in Deliverables 4.1 [MI10] and 4.2 [DDL11], but applied to the biomedical domain. That is to say, while the information in the knowledge infrastructure of the general prototype of WP4 contains ontologies describing common sense knowledge, the knowledge infrastructure for the prototype described in this deliverable contains predominantly ontologies and knowledge sources from the biomedical domain because these ontologies describe segments from this subject domain, and will allow the identification of the entities of interest in the patents. The complete list of the semantic resources that are loaded in the semantic repository is provided in Appendix A.

The architecture of the patent retrieval system is already described in Deliverable 7.1 [CEEB+12]. In order to integrate the information from the processed documents with the knowledge infrastructure, they are indexed and the metadata obtained through their processing are converted into RDF[8], based on the domain specific ontologies, and inserted in the semantic repository (OWLIM [BKO+11]), which stores the knowledge infrastructure, and provides access to the data in it.

Figure 4 illustrates the semantic annotation process. It shows how the words found in the patent text are interpreted as named entities, and how additional information can be obtained about them through the knowledge sources available in the semantic repository. For instance, the word *ampicillin* is recognized as an *FDA Drug* which has dosage forms, and the word *aggression* is recognized as a *disease*.

The recognition of the entities in the texts is performed by a GATE[9] pipeline. Gazetteers (cf. Figure 5) are built to help recognize and annotate the following entities: DiseaseOrDysfunction, AnatomicalStructure, RouteOfAdministration, Drug, ActiveIngredient, DosageForm and Reference. The patents are processed by the tagging tool (GATE v6.1[10]), which add semantic annotations to the words from the patents, cf. Figure 6. The custom configuration of the tool is available in MOLTO repository at:

svn://molto-project.eu/wp7/tools.

The semantic annotation step is followed by a process in which the annotations are extracted from the patents and RDF-ized, i.e., turned into RDF triples. This process unifies their format with the rest of the semantic sources described above. This step is processed with the GateToRdf tool[11]. This tool connects each patent identifier with the annotations that are mentioned in it and with the predicate described in

http://proton.semanticweb.org/protonm#mentions.

Then, this predicate is used during the search phase to select which concepts are mentioned in the document. The GateToRdf tool takes 2 parameters: InputFolderName and OutputFileName. The input folder name is the name of the folder that contains annotated

[8] http://www.w3.org/RDF/

[9] http://gate.ac.uk/

[10] http://gate.ac.uk/

[11] svn://molto-project.eu/wp7/tools/GateToRDF

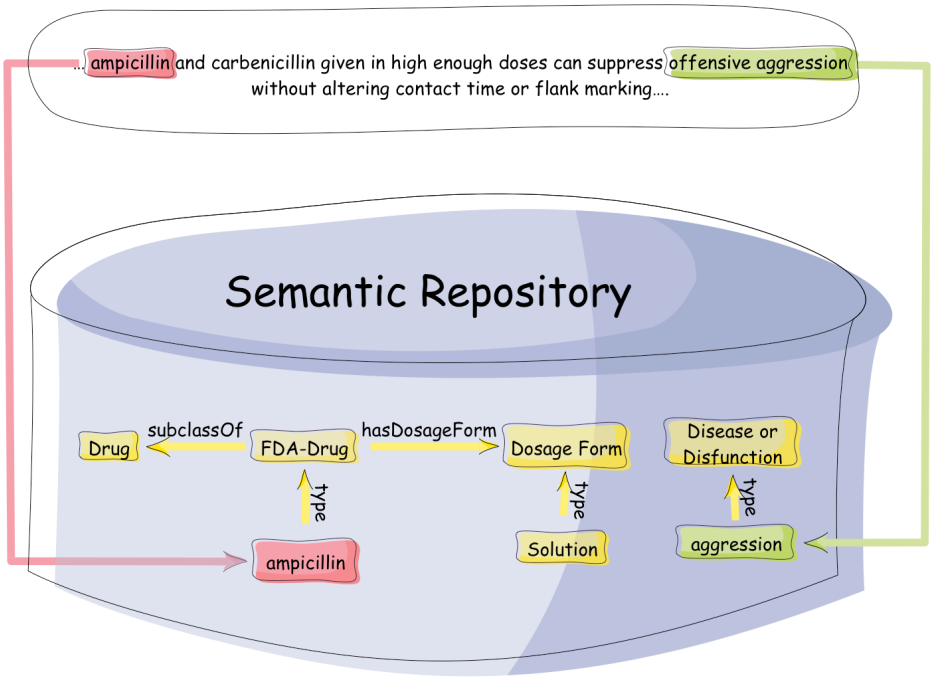

Figure 4: Semantic annotation process

```
FDA Drug
    ampicillin
    cephalothin sodium
    genomil
    penicilline
    permapen
    ...
```

Figure 5: Gazetteer examples being entities of *FDA-Drug*

patents; the outputFileName is the name of the file in which the extracted triples will be stored.

Consequently, the RDF triples are loaded and stored in the semantic repository (OWLIM). This allows to obtain information regarding both patent documents and the characteristics of the drugs, diseases and other entities of interest available in the semantic knowledge base. For instance, the query "information about Ampicillin", which can be run in the online interface, shows the results coming from the documents and from the knowledge bases.

2.4 Query Grammars

The query grammars have been refactored using the set of primitives defined in the Query Library work conducted in WP4. The main purpose of the GF Query Library is to obtain

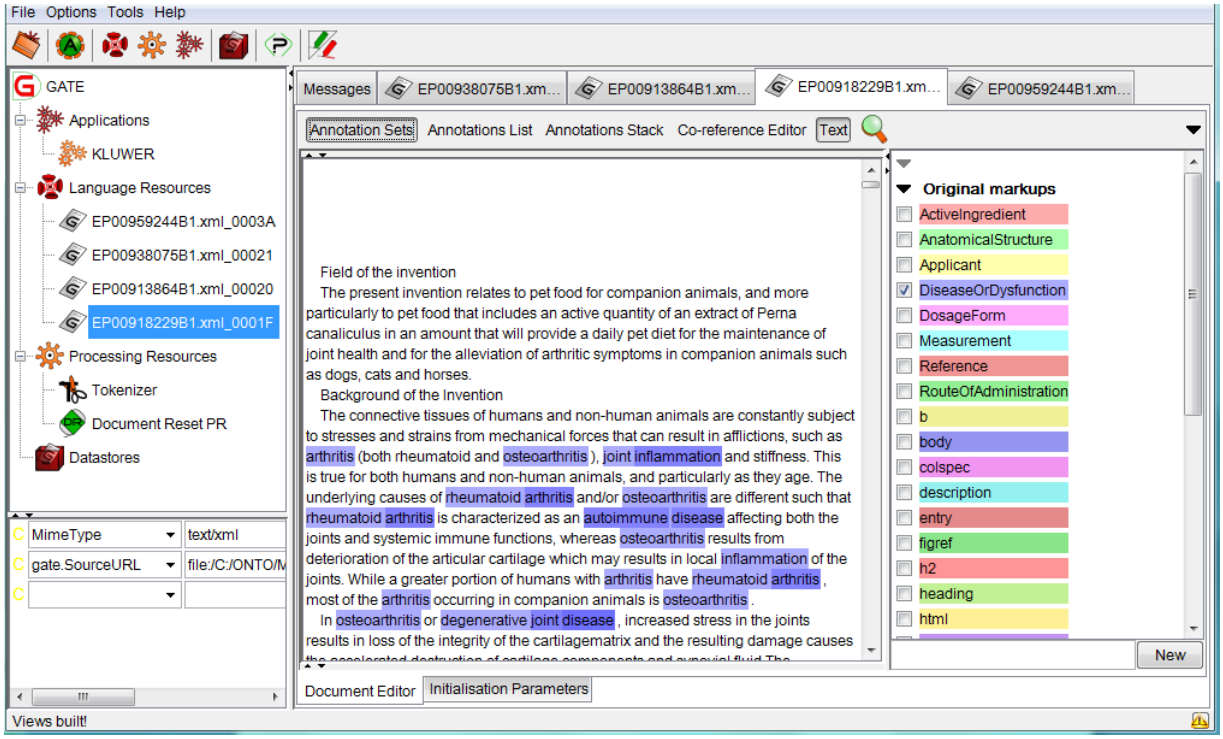

Figure 6: Annotated text in Gate

an unified query grammar that can be used for multiple domains and then specialized according to the specific needs [DDL11, CRDE12]. In consequence, UGOT has adapted the English and French version of the patents query grammar to the new structure, and the German version has been developed from scratch.

From the functionality and coverage point of view, the new grammar is equivalent to the old one. The difference however is the fact that it relies on the primitive query building functions defined in the Query Library. For this reason, the grammar developed for the patents prototype represents a good showcase for the Query Library, showing that it is a valuable resource for writing query grammars for various domains[12] and in a number of languages (English, French, German, Swedish, Bulgarian). Another advantage is that the grammars are easier to test and debug, since they rely on the primitives from the Query Library, which were tested for a number of grammars before. In addition to this, developing a grammar using the Query Library requires less linguistic knowledge, but just selecting the right set of primitives that would be right for the task. An important remark is that the refactoring of the query grammar only refers to the concrete syntaxes, because

[12] So far, the query grammar have been integrated with the patent ontologies (WP7) and the upper level PROTON ontology (WP4), although PROTON ontology is being used in the cultural heritage domain (WP8) as well. Nonetheles, Ontotext have a collection of RDF stores were to apply the query library, like http://linkedlifedata.com and http://ff-dev.ontotext.com that is a collection of common sense knowledge domain of Linked Open Data, sports, news, architecture and food recommendation domains to name a few.

the abstract syntax is still the same, since it refers to new categories and functions specific to the patent domain.

In comparison to the previous patent query grammar, now it has fewer constructions, because of the fact that it is developed on top of the Query Library. As a consequence, the constructions are also more natural and the number of malformed constructions have decreased considerably. The current grammar consists of 31 patterns and it is able to parse/-generate 359 query constructions in English, 111 in French and 147 in German. However, the situation might change after evaluating the Query Library and the two grammars build on top of it and decide upon extending it or restricting certain constructions. The up to date complete list of query topics, patterns and some construction examples can be seen in the Appendix B.

In the following example, the function *PQActive* is used to *ask about the active ingredients of a drug*. Note that the number of alternatives depends on the verbosity and the fertility of the grammar and the rule:

```
PQActive : Drug -> Query ;
```

The *PQActive* function produces the alternative formulations shown in Table 3.

English:
give me all information about all active ingredients of DRUG
all information about all active ingredients of DRUG
give me all information about the active ingredients of DRUG
all information about the active ingredients of DRUG
active ingredients of DRUG
all active ingredients of DRUG
the active ingredients of DRUG
French:
montrer toutes les informations sur tous les ingrèdients actifs de DRUG
des ingrèdients actifs de DRUG
tous les ingrèdients actifs de DRUG
German:
zeigen Sie alle Informationen über alle aktiven Zutaten von DRUG
aktive Zutaten von DRUG
alle aktiven Zutaten von DRUG

Table 3: Alternative formulations for function PQActive

The English concrete syntax for the function had the following form in the previous version of the grammar. It can be noticed the need to use basic syntactic primitives, such as predication and complementation, and also a more low-level manipulation of the GF resource grammar library functions.

```
PQActive drug =
   let
          ai : CN = mkCN active _ingredient _CN (Syntax.mkAdv possess _Prep drug) ;
          sg _df : NP = (mkNP the _Art NumPl ai)   mkNP all _Predet (mkNP the _Art NumPl ai) ;
          massdf : NP = massInfoPl ai
   in
     mkUtt (mkQCl whatPl _IP (mkVP sg _df))
     mkUtt massdf
     mkUtt sg _df
     mkUtt (mkImp (giveMe sg _df))
     mkUtt  (mkCl (mkNP i _Pron) (mkVP (mkVPSlash want _V2) sg _df))  ;
```

The new version of the same function alleviates over these problems by building the same sentences as combination of primitives from the Query Library, which in turn, use the GF resource grammar library primitives. This layering reduces the need for linguistic skills, making it easier for a larger category of users to build their own query grammars.

```
PQActive drug =
   let
        ai : Kind = KRelSet active _ingredient _CN (DrugToSet drug) ;
        sg _df : Set = SAll df ;
        massdf : Set = SPlural df
   in
     QInfo sg _df
     QMass massdf
     QMass sg _df ;
```

However, the new approach does not completely reduce the need for writing queries from scratch, as there could be cases when very specific and idiomatic constructions are not covered by the basic library. However, for the most common ways of expressing a query, assembling the primitives from the basic library should be enough. Indeed, the Query Library was extended with some of the patent constructions since they all had common sense. The fact is that the Query Library is not meant to be an exhaustive collection of patterns, so if a common-sense example appears, one can always extend the library with a new instance.

The resources developed for the patents use case are available at the MOLTO repository[13]. The main grammar files are QueryPats.gf - abstract syntax and QueryPatsEng.gf, QueryPatsFre.gf, QueryPatsGer.gf - concrete grammars for English, French and German. The Query Library is located in the same repository[14]. The main files are named Query.gf - abstract syntax and QueryEng.gf, QueryFre.gf and QueryGer.gf - concrete syntaxes for the above-mentioned languages. In order to compile the grammar one needs to have GF installed, as well as the GF resource grammar library. Consequently one can compile the grammars using the makefile from:

[13] svn://molto-project.eu/wp7/query/patents

[14] svn://molto-project.eu/wp7/query/

svn://molto-project.eu/wp7/query/
or the command gf -make QueryPatsEng.gf QueryPatsFre.gf QueryPatsGer.gf in:
svn://molto-project.eu/wp7/query/patents.

3 The Online Interface to access the Patents Retrieval Prototype

As previously mentioned, the retrieval system can be accessed online at:
http://molto-patents.ontotext.com.

The general walkthrough for the application is shown in Figure 7. The interface allows for querying the system in the EPO official languages, i.e., English, German and French, and the queries are written using the controlled natural language described by the patents query grammar, as seen in Section 2.4. The GF engine gives the abstract representation of the user's query and the retrieval system converts it into SPARQL in order to use it to search for the domain concepts and the documents that are related to the query criteria. The results obtained, i.e., the list of domain concepts and documents, are displayed in an interactive graphical interface that allows for browsing the ontology and inspecting the patent documents.

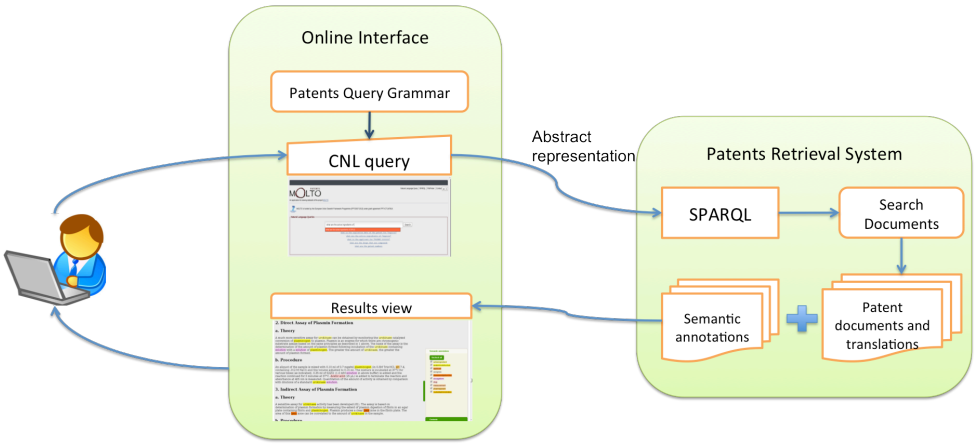

Figure 7: The online interface architecture of the patents prototype

In the previous version of the demo, the query interface presented several natural language query examples, with the purpose of assisting novel users with formulating his or her first requests to the system. However, several of the examples did not return any supporting patents containing the information requested by the query. Even trying out many queries in a row may result in answers that are not supported by patents from the collection. In order to overcome such frustrating attempts, we took two approaches. First, we changed the example queries from the demo page in order to ensure that they do return some supporting patents. Second, in the deliverable we present a summary (roadmap) of

the relations present in the ontology that are also supported by patents, such that users that want to test our system can have a comprehensive set of examples to start with. In what follows we give more details on each.

3.1 NL query examples

The demo interface allows the user to give the search criteria using a controlled natural language. Every possible user's input described by the controlled language has a correspondence with a grammar pattern in the grammar, and it generates an abstract syntax tree that is translated into SPARQL[15]. Besides, the query grammar has been integrated in the interface in order to enable an autocomplete function to help the user writing queries under the controlled language, as shown in Figure 8.

The new queries that are given as examples on the interface are:

give me all information about AMPICILLIN	12 documents
give me all information about all active ingredients of BACLOFEN	21 documents
give me all information about all routes of administration of FAMOTIDINE	24 documents
give me all information about all dosage forms of GANCICLOVIR	24 documents
give me the approval date of the patent for REBETOL	6 documents

Table 4: Query Examples and the number of results.

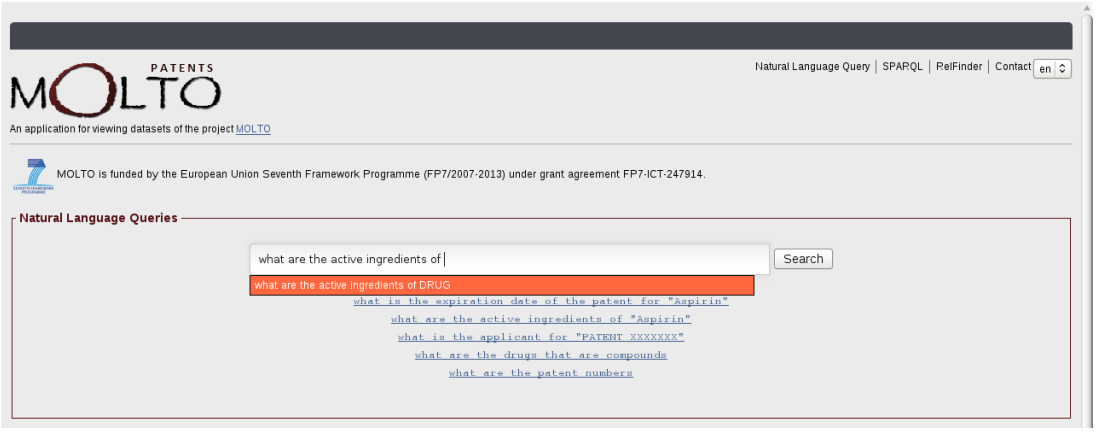

Figure 8: The natural language query interface

3.2 Database Roadmap

Below we present for each type of natural-language query a general table, containing entity names that can be used in order to obtain non-empty results.

[15]http://www.w3.org/TR/rdf-sparql-query/

Query type: `Give me all information about *drug/active ingredient*'

Table 9 shows drugs that are involved in triples in the ontology and are also mentioned in patents. If the *drug* is chosen among the drugs in the table, then the query `Give me all information about *drug*' is guaranteed to return supporting patents. Table 9 is abbreviated, for a complete table run the following general SPARQL query:

```
SELECT ?drug (count(distinct ?doc) as ?count)
WHERE
    ?s ?p ?o .
    ?s <http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#hasName> ?n .
    ?n <http://www.w3.org/1999/02/22 rdf syntax ns#type>
       <http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#FDA_DrugName> .
    ?n <http://www.w3.org/2000/01/rdf schema#label> ?drug .

   UNION
    ?o ?p ?s .
    ?s <http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#hasName> ?n .
    ?n <http://www.w3.org/1999/02/22 rdf syntax ns#type>
       <http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#FDA_DrugName> .
    ?n <http://www.w3.org/2000/01/rdf schema#label> ?drug .

   ?doc <http://proton.semanticweb.org/protonm#mentions> ?n .

GROUP BY ?drug
ORDER BY desc(?count)
```

A similar query for obtaining a summary of the active ingredients mentioned in the collection of patents can be obtained by the following query:

```
SELECT ?ai (count(distinct ?doc) as ?count)
WHERE
    ?s ?p ?n .
    ?n <http://www.w3.org/1999/02/22 rdf syntax ns#type>
       <http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#ActiveIngredient
           > .
    ?n <http://proton.semanticweb.org/protonsys#mainLabel> ?ai .

   UNION
    ?n ?p ?s .
    ?n <http://www.w3.org/1999/02/22 rdf syntax ns#type>
       <http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#ActiveIngredient
           > .
    ?n <http://proton.semanticweb.org/protonsys#mainLabel> ?ai .

   ?doc <http://proton.semanticweb.org/protonm#mentions> ?n .

GROUP BY ?ai
ORDER BY desc(?count)
```

Query type: `Give me all information about the active ingredients of *drug*'

Table 11 shows a small fraction of the drugs for which active ingredients are known (present in the ontology via the relation hasActiveIngredient) and these ingredients are mentioned in patent documents. The full list can be obtained with following SPARQL query, via the SPARQL interface of the demo. All queries including drugs listed in Table 11 are guaranteed to return supporting documents.

```
SELECT ?drug ?l (count(distinct ?doc) as ?count)
WHERE
    ?o <http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#
        hasActiveIngredient> ?d .
    ?o <http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#hasName> ?s .
    ?s <http://www.w3.org/1999/02/22 rdf syntax ns#type>
       <http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#FDA_DrugName> .
    ?s <http://www.w3.org/2000/01/rdf schema#label> ?drug .
    ?d <http://www.w3.org/2000/01/rdf schema#label> ?l .
    ?doc <http://proton.semanticweb.org/protonm#mentions> ?d .
    ?doc <http://proton.semanticweb.org/protonm#mentions> ?s .

GROUP BY ?drug ?l
ORDER BY ?drug
```

Query type: `Give me all information about all routes of administration of *drug*'

A table with drugs, routes of administrations and number of documents mentioning the drug and the route of administrations can be obtained via the query:

```
SELECT ?label ?l (count(distinct ?doc) as ?count)
WHERE
  ?s <http://www.w3.org/2000/01/rdf schema#label> ?label .
  ?s <http://www.w3.org/1999/02/22 rdf syntax ns#type>
     <http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#FDA_DrugName> .
  ?o <http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#hasName> ?s.
  ?o <http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#
      hasRouteOfAdministration> ?d.
  ?d <http://www.w3.org/2000/01/rdf schema#label> ?l.

  ?doc <http://proton.semanticweb.org/protonm#mentions> ?d.
  ?doc <http://proton.semanticweb.org/protonm#mentions> ?s

GROUP BY ?label ?l
ORDER BY desc(?count)
```

Table 12 shows a part of the results returned by the query above. If the user chooses one of the drugs from the table for a query of the type `Give me all information about all routes of administration of *drug*', then there will be documents returned.

Query type: `give me all information about all dosage forms of *drug*'

A table of all drugs, together with their dosage forms and the number of documents that contain related in formation can be obtained by running the following general SPARQL query:

```
SELECT ?label ?l (count(distinct ?doc) as ?count)
WHERE
  ?s <http://www.w3.org/2000/01/rdf schema#label> ?label .
  ?s <http://www.w3.org/1999/02/22 rdf syntax ns#type>
     <http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#FDA_DrugName> .
  ?o <http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#hasName> ?s.
  ?o <http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#hasDosageForm> ?d.
  ?d <http://www.w3.org/2000/01/rdf schema#label> ?l.

  ?doc <http://proton.semanticweb.org/protonm#mentions> ?d.
  ?doc <http://proton.semanticweb.org/protonm#mentions> ?s
```

```
GROUP BY ?label ?l
ORDER BY desc(?count)
```

3.3 Queries interpretation

In the earlier versions of the demo, the query type 'give me all information about all routes of administration of *drug*' was returning confusing results. Specifically, the query was interpreted as follows: `find in the ontology the routes of administration of *drug*, then search for documents mentioning these routes of administration (independently from the drug)'. As a result, the user obtains for example a long list of patents mentioning various drugs that are administered orally (if *drug* is administered orally). The meaning of the initial natural-language query is different though, the user searching for information on how the *drug* can be administered.

We corrected the interpretation of the query and correspondingly, the SPARQL translation, as follows. In the new query, documents that mention both the *drug* and the route of administration must be mentioned by the document. Below is the updated query:

```
CONSTRUCT
  ?d <http://www.w3.org/2000/01/rdf schema#label> ?l.
  ?doc <http://proton.semanticweb.org/protonm#mentions> ?d

WHERE
  ?s <http://www.w3.org/2000/01/rdf schema#label> "FAMOTIDINE" .
  ?s <http://www.w3.org/1999/02/22 rdf syntax ns#type>
     <http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#FDA_DrugName> .
  ?o <http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#hasName> ?s.
  ?o <http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#
      hasRouteOfAdministration> ?d.
  ?d <http://www.w3.org/2000/01/rdf schema#label> ?l.
  OPTIONAL
      ?doc <http://proton.semanticweb.org/protonm#mentions> ?d.
      ?doc <http://proton.semanticweb.org/protonm#mentions> ?s
```

However, the co-occurence of the two entities in the document does not guarantee that the document describes precisely the drug and its route of administration. For example, if the target drug is **D** and its route of administration is **oral**, and **G** is another drug, then a document containing the following sentence will be returned by the query: `Drug **G** is administered **orally**. If taken at the same time with drug **D**, drug **G** will lead to severe side effects.' This document will be returned by the query, although there is no mention of how drug **D** is administered.

In order to avoid such false results, more complex approaches are necessary, for example methods for automated extracting of relations between annotated entities in text. Such methods exist [MPK+05] and can be considered for future developments of the project.

References

[BKO+11] Barry Bishop, Atanas Kiryakov, Damyan Ognyanoff, Ivan Peikov, Zdravko Tashev, and Ruslan Velkov. OWLIM: A family of scalable semantic repositories. *Semantic Web Journal*, 2:33--42, June 2011.

[CEEB+12] Milen Chechev, Ramona Enache, Cristina España-Bonet, Meritxell Gonzalez, Lluís Marquez, Borislav Popov, and Aarne Ranta. D7.1. Patent MT and Retrieval Prototype Beta, January 2012.

[CRDE12] Milen Chechev, Aarne Ranta, Mariana Damova, and Ramona Enache. Grammar Ontology Interoperability, May 2012.

[DDL11] Mariana Damova, Dana Dannélls, and Inari Listenmaa. D4.2 Data Models and Alignments, May 2011.

[EBGM11] Cristina España-Bonet, Meritxell Gonzalez, and Lluís Marquez. D5.1 Description of the final collection of corpora, September 2011.

[KHM+07] Philipp Koehn, Hieu Hoang, Alexandra Birch Mayne, Christopher Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open Source Toolkit for Statistical Machine Translation. In *Annual Meeting of the Association for Computation Linguistics (ACL), Demonstration Session*, pages 177--180, Jun 2007.

[KSF+06] Philipp Koehn, Wade Shen, Marcello Federico, Nicola Bertoldi, Chris Callison-Burch, Brooke Cowan, Chris Dyer, Hieu Hoang, Ondrej Bojar, Richard Zens, Alexandra Constantin, Evan Herbst, and Christine Moran. Open Source Toolkit for Statistical Machine Translation. Technical report, Johns Hopkins University Summer Workshop. http://www.statmt.org/jhuws/, 2006.

[MI10] Petar Mitankin and Atanas Ilchev. Knowledge Representation Infrastucture, 2010.

[MOL12] MOLTO. D5.2. Description and Evaluation of the Combination Prototypes , March 2012.

[MPK+05] Ryan McDonald, Fernando Pereira, Seth Kulick, Scott Winters, Yang Jin, and Pete White. Simple algorithms for complex relation extraction with applications to biomedical ie. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 491--498, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[Och03] Franz Josef Och. Minimum error rate training in statistical machine translation. In *Proc. of the Association for Computational Linguistics*, Sapporo, Japan, July 6-7 2003.

[ON03] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19--51, 2003.

[PRWZ02] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the Association of Computational Linguistics*, pages 311--318, 2002.

[Sto02] A. Stolcke. SRILM -- An extensible language modeling toolkit. In *Proc. Intl. Conf. on Spoken Language Processing*, 2002.

A Ontologies in the Biomedical Domain

For the patent retrieval prototype several ontologies from the biomedical domain are used. This is the complete list of all datasets that are loaded and the description of their content.

kb/FDA/FDA_classonly_2.owl: FDA Products naive ontology created and aligned with a basic upper level ontology PROTON; (The diagram is already attached to previous deliverable).

FDA_to_KIM.nt: mapping between FDA_classonly_2 and proton classes.

FDA_products.nt: triples extracted with gazetteers from FDA Orange Book.

measure-unit-classes.owl: ontology with measurement units.

measure-unit-instances.owl: instances of measurements.

unit-main-labels.nt: labels of the measurement units.

proton-measure.owl: mapping between proton and measurements ontology.

proton-patents.owl: patent structure ontology.

skos.rdf: skos ontology(Simple Knowledge Organization System).

umls-semnet-proton.nt: umls general concepts.

pathologic-functions.nt: umls pathologic functions instances.

anatomical-structures.nt: umls anatomical structures instances.

umls-gpcr-proteins.nt: umls gpcr proteins instances.

pharma-params-instances.nt: define parameters to.

owl.rdfs: specifies in RDF Schema format the built-in classes and properties that together form the basis of the RDF/XML syntax of OWL Full, OWL DL and OWL Lite.

protonsys.n3;protontop.n3;protonext.n3;protonkm.n3: the Proton upper level ontology.

annotated_docs_tripples.n3: annotations from currently used patents.

B Topics, Patterns and Constructions for the Patents Query Library

The patents query grammar covers a set of query topics, shown in Table 5. We wrote a number of query examples for each topic, and from those examples we wrote pattern rules for the patent query grammar. The current grammar consists of the 31 patterns shown and it generates 359 query constructions in English, 111 in French and 147 in German. Tables 6,7 and 8 show some examples of the patent queries in the three languages.

information about a drug	drugs that are compounds
active ingredients of a drug	drug preparations
dosage forms of a drug	the name of a drug
route of administration of a drug	methods in the patent
dosage form of a drug	use of patent
patent number	use of drug
the expiration of a patent	strength of a drug
patent use codes	claims from a date that mention a given drug
patent application number	claims about a given drug authored by somebody
applicant for a patent	approval date of a patent

Table 5: The patent query topics

Pattern: PQActive Aspirin
give me all information about all active ingredients of DRUG
all information about all active ingredients of DRUG
give me all information about the active ingredients of DRUG
all information about the active ingredients of DRUG
active ingredients of DRUG
all active ingredients of DRUG
the active ingredients of DRUG
Pattern: PQCompounds
give me all information about all drugs that are compounds
all information about all drugs that are compounds
give me all information about the drugs that are compounds
all information about the drugs that are compounds
drugs that are compounds
all drugs that are compounds
the drugs that are compounds
Pattern: PQDrugPrep Aspirin
QueryPatsEng: give me the drug preparation for DRUG
give me the names of the drug preparation for DRUG
what is the drug preparation for DRUG
what are the names of the drug preparation for DRUG
which is the drug preparation for DRUG
which are the names of the drug preparation for DRUG
the drug preparation for DRUG
the names of the drug preparation for DRUG
drug preparation for DRUG

Table 6: The patent query examples in English

Pattern: PQActive Aspirin
zeigen Sie alle Informationen lle aktiven Zutaten von DRUG
aktive Zutaten von DRUG
alle aktiven Zutaten von DRUG
PQCompounds
zeigen Sie alle Informationen lle Medikamente die Verbindungen sind
Medikamente die Verbindungen sind
alle Medikamente die Verbindungen sind
Pattern: PQDrugPrep Aspirin
zeigen Sie die Medikamentenherstellung für DRUG
Medikamentenherstellung für DRUG
die Medikamentenherstellung für DRUG

Table 7: The patent query examples in German

Pattern: PQActive Aspirin
montrer toutes les informations sur tous les ingrédients actifs de DRUG
des ingrédients actifs de DRUG
tous les ingrédients actifs de DRUG
Pattern: PQCompounds
montrer toutes les informations sur tous les médicaments qui sont des composés
des médicaments qui sont des composés
tous les médicaments qui sont des composés
Pattern: PQDrugPrep Aspirin
montrer la préparation pour DRUG
de la préparation pour DRUG
la préparation pour DRUG

Table 8: The patent query examples in French

C Patent Retrieval Databases Roadmap

The following tables contain part of content of the retrieval system databases. They mostly constitute the concepts contained in the ontologies that can be also find in the patent documents indexed in the system.

Drug	**#docs**	**Drug**	**#docs**
ACETIC ACID	1050	THIOGUANINE	135
SODIUM CHLORIDE	811	VITAMIN D	130
TALC	749	POTASSIUM CHLORIDE	130
INSULIN	711	SIMVASTATIN	128
LENTE	497	KANAMYCIN	128
PENICILLIN	414	HYDROXYUREA	127
SODIUM BICARBONATE	402	NAPROXEN	126
CISPLATIN	319	TESTOSTERONE	125
MAGNESIUM SULFATE	315	FLUTAMIDE	121
CYCLOPHOSPHAMIDE	297	LIDOCAINE	119
ADENOSINE	294	AZATHIOPRINE	118
FLUOROURACIL	288	THIOTEPA	117
DIMETHYL SULFOXIDE	277	VITAMIN A	115
AMMONIUM CHLORIDE	269	DACARBAZINE	115
DEXAMETHASONE	268	TAXOTERE	114
ETOPOSIDE	260	GLUCAGON	110
STERILE WATER	259	HYDROCORTISONE	107
PACLITAXEL	244	PROGESTERONE	106
MITOMYCIN	218	METHYLPREDNISOLONE	104
CARBOPLATIN	208	NICOTINE	98
TAXOL	208	BENZYL BENZOATE	96
CYCLOSPORINE	207	CLADRIBINE	96
AMPICILLIN	203	PENTOSTATIN	95
PREDNISONE	194	PIROXICAM	94
IBUPROFEN	192	RIBAVIRIN	94
MERCAPTOPURINE	170	GENTAMICIN	92
MITOXANTRONE	170	KETOPROFEN	92
ESTRADIOL	170	FLUOXETINE	92
CYTARABINE	159	ACETAMINOPHEN	91
INDOMETHACIN	155	NIFEDIPINE	91
FOLIC ACID	151	TRIAMCINOLONE	91
PREDNISOLONE	151	SODIUM THIOSULFATE	87
IFOSFAMIDE	141	CAPTOPRIL	85
LOVASTATIN	138	...&...	

Table 9: Drugs mentioned in patents.

Active Ingredient	**#docs**	**Active Ingredient**	**#docs**
CALCIUM	1458	AMMONIUM CHLORIDE	269
ALCOHOL	1421	DEXAMETHASONE	268
AMINO ACIDS	1138	ETOPOSIDE	260
GLYCERIN	957	PACLITAXEL	244
SODIUM CHLORIDE	811	MITOMYCIN	218
MANNITOL	806	ASPIRIN	211
TALC	749	TETRACYCLINE	209
GLYCINE	716	CARBOPLATIN	208
CITRIC ACID	663	CYCLOSPORINE	207
SORBITOL	655	PREDNISONE	194
SULFUR	631	IBUPROFEN	192
TYROSINE	595	DOCETAXEL	191
PROTEASE	577	MELPHALAN	182
GLUTAMINE	565	SOYBEAN OIL	181
PHOSPHORIC ACID	530	VITAMIN E	179
LACTIC ACID	465	HYDROXYPROPYL CELLULOSE	179
UREA	452	ISOPROPYL ALCOHOL	177
ASCORBIC ACID	444	MERCAPTOPURINE	170
SODIUM CARBONATE	444	ESTRADIOL	170
TARTARIC ACID	439	CETYL ALCOHOL	168
DEXTROSE	436	CHLORAMBUCIL	164
SODIUM BICARBONATE	402	NITRIC OXIDE	160
BIOTIN	389	CYTARABINE	159
COPPER	381	INDOMETHACIN	155
CALCIUM CARBONATE	379	FOLIC ACID	151
SODIUM SULFATE	346	PREDNISOLONE	151
SODIUM ACETATE	321	BUSULFAN	144
CISPLATIN	319	LIPASE	144
MAGNESIUM SULFATE	315	IFOSFAMIDE	141
SODIUM CITRATE	312	ALUMINUM HYDROXIDE	141
CYCLOPHOSPHAMIDE	297	TENIPOSIDE	140
ADENOSINE	294	LOVASTATIN	138
FLUOROURACIL	288	DACTINOMYCIN	138
SODIUM PHOSPHATE	279	CALCIUM CHLORIDE	136
DIMETHYL SULFOXIDE	277	...&...	

Table 10: Active ingredients mentioned in patents.

Drug	Active Ingredient	#docs
ABILIFY	ARIPIPRAZOLE	3
ABRAXANE	PACLITAXEL	16
ACARBOSE	ACARBOSE	69
ACCOLATE	ZAFIRLUKAST	3
ACCUPRIL	QUINAPRIL HYDROCHLORIDE	3
ACCUTANE	ISOTRETINOIN	4
ACEON	PERINDOPRIL ERBUMINE	5
ACETAMINOPHEN	ACETAMINOPHEN	91
ACETAZOLAMIDE	ACETAZOLAMIDE	35
ACETOHEXAMIDE	ACETOHEXAMIDE	26
ACETYLCYSTEINE	ACETYLCYSTEINE	44
ACTH	CORTICOTROPIN	6
ACTONEL	RISEDRONATE SODIUM	1
ACTRON	KETOPROFEN	1
ACYCLOVIR	ACYCLOVIR	71
ACYCLOVIR	ACYCLOVIR SODIUM	2
ACYCLOVIR SODIUM	ACYCLOVIR SODIUM	2
ADALAT	NIFEDIPINE	3
ADENOSINE	ADENOSINE	294
ADRUCIL	FLUOROURACIL	6
ADVICOR	NIACIN	6
ADVICOR	LOVASTATIN	8
ADVIL	IBUPROFEN	3
AGENERASE	AMPRENAVIR	6
ALA-CORT	HYDROCORTISONE	4
ALBUTEROL	ALBUTEROL	51
ALBUTEROL SULFATE	ALBUTEROL SULFATE	10
ALDACTONE	SPIRONOLACTONE	1
ALDARA	IMIQUIMOD	3
ALENDRONATE SODIUM	ALENDRONATE SODIUM	7
ALEVE	NAPROXEN SODIUM	1
ALIMTA	PEMETREXED DISODIUM	1
ALKERAN	MELPHALAN	7
ALLOPURINOL	ALLOPURINOL	29
ALLOPURINOL SODIUM	ALLOPURINOL SODIUM	2
ALOCRIL	NEDOCROMIL SODIUM	1
ALORA	ESTRADIOL	1
ALOXI	PALONOSETRON HYDROCHLORIDE	5
ALPHADERM	HYDROCORTISONE	1
ALPHAGAN	BRIMONIDINE TARTRATE	2
ALPHAGAN P	BRIMONIDINE TARTRATE	2
ALPRAZOLAM	ALPRAZOLAM	25
ALPROSTADIL	ALPROSTADIL	8
ALTACE	RAMIPRIL	7
AMARYL	GLIMEPIRIDE	2
AMCINONIDE	AMCINONIDE	11
AMIFOSTINE	AMIFOSTINE	28
AMIKACIN SULFATE	AMIKACIN SULFATE	2
AMINOCAPROIC	AMINOCAPROIC ACID	25
AMINOCAPROIC ACID	AMINOCAPROIC ACID	25
AMINOPHYLLIN	AMINOPHYLLINE	7
AMINOPHYLLINE	AMINOPHYLLINE	14
AMITIZA	LUBIPROSTONE	1
AMLEXANOX	AMLEXANOX	4
AMLODIPINE BESYLATE	AMLODIPINE BESYLATE	9
AMMONIUM CHLORIDE	AMMONIUM CHLORIDE	269
AMMONIUM LACTATE	AMMONIUM LACTATE	6
AMOXAPINE	AMOXAPINE	31
AMOXICILLIN	AMOXICILLIN	41
AMOXIL	AMOXICILLIN	1
AMPHOTEC	AMPHOTERICIN B	1
AMPHOTERICIN B	AMPHOTERICIN B	62
...	...	...

Table 11: Drug names, the active ingredients of which are mentioned in documents.

Drug	Administration	#docs
SELENIUM SULFIDE	TOPICAL	2
REBETOL	ORAL	6
PROCHLORPERAZINE	ORAL	21
PROCHLORPERAZINE	RECTAL	15
PROCHLORPERAZINE	INJECTION	21
OCTREOTIDE ACETATE	INJECTION	6
MISOPROSTOL	ORAL	17
ALBUTEROL SULFATE	ORAL	10
ALBUTEROL SULFATE	INHALATION	7
GENTAMICIN	TOPICAL	47
GENTAMICIN	INJECTION	64
NALBUPHINE	INJECTION	16
ETHOSUXIMIDE	ORAL	20
POTASSIUM CITRATE	ORAL	13
PYRILAMINE MALEATE	ORAL	5
HYZAAR	ORAL	6
NIZATIDINE	ORAL	6
VEPESID	ORAL	6
VEPESID	INJECTION	6
NEOSAR	INJECTION	5
TARCEVA	ORAL	44
MERCAPTOPURINE	ORAL	159
CARBAMAZEPINE	ORAL	68
RISPERDAL	ORAL	7
CLOTRIMAZOLE	ORAL	30
CLOTRIMAZOLE	TOPICAL	23
CLOTRIMAZOLE	VAGINAL	14
SUSTIVA	ORAL	7
SECOBARBITAL SODIUM	INJECTION	2
SECOBARBITAL SODIUM	ORAL	2
PERPHENAZINE	ORAL	20
INDOCIN	ORAL	5
INDOCIN	RECTAL	2
CHANTIX	ORAL	2
ALEVE	ORAL	2
VERAPAMIL HYDROCHLORIDE	ORAL	4
VERAPAMIL HYDROCHLORIDE	INJECTION	3
GUANFACINE HYDROCHLORIDE	ORAL	1
DEPO-PROVERA	INJECTION	1
STAPHCILLIN	INJECTION	1
TRILAFON	ORAL	1
TRILAFON	INJECTION	1
ECONAZOLE NITRATE	TOPICAL	4
TRICOR	ORAL	7
PROVERA	ORAL	3
DILAUDID	ORAL	1
PARACORT	ORAL	8
LOGEN	ORAL	13
CIPROFLOXACIN	OPHTHALMIC	17
CIPROFLOXACIN	INJECTION	40
CHOLESTYRAMINE	ORAL	50
CONCERTA	ORAL	1
WARFARIN SODIUM	ORAL	6
VITAMIN D	ORAL	100
LYRICA	ORAL	1
AMOXAPINE	ORAL	29
PANRETIN	TOPICAL	3
SYMLIN	SUBCUTANEOUS	2
FLUMAZENIL	INJECTION	12
DRONABINOL	ORAL	32
DIPROSONE	TOPICAL	1
METICORTEN	ORAL	4
...	...	...

Table 12: Drug names and routes of administration that are mentioned in documents.

Drug	Dosage Form	#docs	Drug	Dosage Form	#docs
TALC	POWDER	570	ESTRADIOL	TABLET	78
SODIUM CHLORIDE	INJECTABLE	429	IBUPROFEN	TABLET	77
DIMETHYL SULFOXIDE	SOLUTION	276	IFOSFAMIDE	INJECTABLE	76
INSULIN	INJECTABLE	265	INDOMETHACIN	CAPSULE	75
FLUOROURACIL	SOLUTION	256	AMPICILLIN	CAPSULE	72
DEXAMETHASONE	SOLUTION	247	BALANCED SALT	SOLUTION	71
LENTE	INJECTABLE	242	PREDNISONE	TABLET	70
STERILE WATER	LIQUID	229	ISOFLURANE	LIQUID	70
SODIUM BICARBONATE	INJECTABLE	183	POTASSIUM IODIDE	SOLUTION	70
CYCLOSPORINE	SOLUTION	181	MEGESTROL ACETATE	SUSPENSION	66
PREDNISONE	SOLUTION	176	THIOTEPA	INJECTABLE	65
FLUOROURACIL	INJECTABLE	153	SIMVASTATIN	TABLET	65
IBUPROFEN	SUSPENSION	150	LOVASTATIN	TABLET	64
CISPLATIN	INJECTABLE	149	MERCAPTOPURINE	TABLET	62
DEXAMETHASONE	INJECTABLE	147	HYDROCORTISONE	POWDER	62
CYCLOPHOSPHAMIDE	INJECTABLE	140	TAXOTERE	INJECTABLE	61
MAGNESIUM SULFATE	INJECTABLE	133	DIAZEPAM	SOLUTION	60
ETOPOSIDE	INJECTABLE	133	DACARBAZINE	INJECTABLE	59
AMMONIUM CHLORIDE	INJECTABLE	127	FLUOROURACIL	CREAM	58
PACLITAXEL	INJECTABLE	126	TESTOSTERONE	INJECTABLE	57
INDOMETHACIN	SUSPENSION	118	THEOPHYLLINE	SOLUTION	56
TAXOL	INJECTABLE	113	FUROSEMIDE	SOLUTION	56
ADENOSINE	INJECTABLE	112	FLUTAMIDE	CAPSULE	55
MITOMYCIN	INJECTABLE	109	HYDROCHLOROTHIAZIDE	SOLUTION	55
CYCLOPHOSPHAMIDE	TABLET	107	POTASSIUM CHLORIDE	INJECTABLE	55
CARBOPLATIN	INJECTABLE	102	PREDNISOLONE	TABLET	55
MITOXANTRONE	INJECTABLE	101	PENTOSTATIN	INJECTABLE	54
CYCLOSPORINE	INJECTABLE	97	TRETINOIN	SOLUTION	53
CYTARABINE	INJECTABLE	96	FLUOXETINE	CAPSULE	53
ETOPOSIDE	CAPSULE	94	FOLIC ACID	INJECTABLE	53
NAPROXEN	SUSPENSION	92	ACYCLOVIR	SUSPENSION	53
IBUPROFEN	CAPSULE	92	CLADRIBINE	INJECTABLE	52
DEXAMETHASONE	TABLET	88	HYDROXYUREA	CAPSULE	51
ERYTHROMYCIN	SOLUTION	80	BENZYL BENZOATE	EMULSION	50
CYCLOSPORINE	CAPSULE	79	...	...	...

Table 13: Drug names and dosage forms that are mentioned in documents.