



Multilingual Online Translation

Non multa, sed multum

D7.3 Patent MT and Retrieval. Final Report.

Contract No.:	FP7-ICT-247914
Project full title:	MOLTO - Multilingual Online Translation
Deliverable:	D7.3 Patent MT and Retrieval. Final Report.
Security (distribution level):	Public, regular publication
Contractual date of delivery:	M33
Actual date of delivery:	March 2013
Type:	Regular Publication
Status & version:	final
Author(s):	Maria Mateva ² , Meritxell Gonzàlez ¹ , Ramona Enache ³ , Cristina España-Bonet ¹ , Lluís Màrquez ¹ , Borislav Popov ² , Aarne Ranta ³
Task responsible:	UPC ¹
Other contributors:	Ontotext ² , UGOT ³

ABSTRACT

The present document is the final report of the Patents Case Study. It describes the multilingual patents retrieval prototype, produced in this work-package, and a brief user manual to access the on-line demonstrative application. The “WP7: Patents Case Study” aims to create a prototype for automatic translation and multilingual retrieval of patents. The main highlights achieved in the final prototype, with respect to the previous versions described in the Deliverable 7.1 [MOL12c] and Deliverable 7.2 [MOL12d], are 1) the translation of patent documents semantically annotated, driven by a statistical machine translation system; 2) a new querying approach for (controlled) natural language to SPARQL translation, driven by the Grammatical Framework and the query grammar on the biomedical domain; 3) the integration of a mechanism for query back-off, based on free text search; 4) some updates on the retrieval system to improve the response time of the prototype, and also 5) some updates on the on-line user interface that address usability aspects and additional functionalities.

Contents

1	Introduction	4
1.1	Motivation	4
1.2	Related Work	4
1.3	The MOLTO approach	5
2	The Patents Case Study Prototype	7
2.1	Patent corpus and translation	8
2.1.1	The biomedical patent datasets	8
2.1.2	Translation of patent documents	9
2.1.3	The patent translation API	11
2.1.4	Syntactic measures for MT evaluation	12
2.2	Patents retrieval system, ontologies and semantic annotations	13
2.2.1	Free text search	15
2.2.2	Ontologies for the biomedical domain	16
2.2.3	Semantic annotations extracted from the patents metadata	17
2.2.4	Semantic annotation over millions of patents	18
2.2.5	Retrieval speed optimization	18
2.2.6	Evaluation of the retrieval system	19
2.3	GF grammars for generation of SPARQL	20
2.3.1	The patents query language	22
2.3.2	Better coverage of the ontology by GF	23
3	The On-line User Interface for the Patents Retrieval Prototype	25
3.1	GUI updates	25
3.1.1	Multilingual annotations	26
3.1.2	Document source exposed	26
3.2	NL query examples	26
3.3	User Scenarios	27
3.3.1	Query types	28
3.3.2	Semantic data exploration	29
3.3.3	Single word query	30
3.3.4	FTS queries	31
4	Conclusions and Challenges	33
A	Example Showing the Translation Steps of an Excerpt of Patent	36
B	Patent Translator API - Specification	38
B.1	Translation of <i>raw</i> text	38
B.2	Translation of <i>XML</i> patent documents	39
C	Ontologies in the Biomedical Domain	40

D Semantic Annotations - Final Types	41
E Biomedical Patents - Query Patterns	42
F Patent Retrieval Databases Roadmap	45

1 Introduction

This document is the final Report of WP7: “Patents Case Study”. This case study aims to create a prototype (the patents 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.

1.1 Motivation

The five major international patent offices in the world maintain patent databases written mainly in English, French, German, Chinese, Japanese and Korean. In addition, many other smaller offices maintain their databases as well, having documents written in their official languages. These offices and their users have the clear need to exchange the content of their databases and the information related to such inventions. This need has actually promoted the organization of international conferences and competitions related to the field, such as patent classification, retrieval and translation, and the development of systems able to search, access and translate patent contents, either from mono-lingual or cross-lingual databases and make them available to the community, ideally in the user’s language.

1.2 Related Work

Most of the public search interfaces, either from the patent offices as the European Patent Office (EPO¹), or independent ones as PatentLens², offer keyword-based search on the title, or multifaceted searching and browsing through the application number, the applicants, inventors and classification codes. Also, systems as Google Patents allow free text search through the original text of the patents.

More recent systems introduced the use of ontologies, which provide a shallow representation of the information space. These light-weight ontologies provide controlled lexicons for the classification of the content. However, few systems take a real advantage of the full potential of an ontological representation. This is the case of the ontology-based retrieval model described in [VFC05]. Their model supports semantic search in document repositories through the exploitation of full-fledged domain ontologies and knowledge bases. As in our system, full documents are returned in response to the user needs. The search system takes advantage of both detailed instance-level knowledge available in the knowledge base, and topic taxonomies for classification.

The BioPatentMiner [MB04] is a more recent system for biomedical information retrieval especially designed to discover relationships among the concepts in the knowledge resources, using a concrete methodology to determine the semantic associations between two resources. The system integrates the patent information from the United States Patent and Trademark Office³ and creates a Biomedical Semantic Web.

¹<http://worldwide.espacenet.com/>

²<http://www.patentlens.net>

³<http://www.uspto.gov>

Regarding patent translation systems, the EPO public service, in combination with the Google Patent Translate service, offers automatic translation of abstracts, descriptions and claims excerpts selected by the user to 14 languages.

A different approach is that of the Patent Language Translations Online (PLuTO) project, a dedicated project to patent translation. Its MT framework is a web service whereby users can request translations. The translation engine uses the MaTrEx (Machine Translation Using Examples) system developed at DCU [AFG⁺06]. It is a hybrid data-driven system built following established design patterns, with an extensible framework allowing for the interchange of novel or previously developed modules as it is defined in [TWS10].

1.3 The MOLTO approach

This report gives an overview of the translation and multilingual retrieval system for biomedical patents developed in the MOLTO, and the resources used to this end, with especial focus on the latest developments.

The patents prototype is publicly available at: <http://molto-patents.ontotext.com/>. It integrates four main components, machine translation, semantic annotation, document retrieval and the on-line user interface. Two different approaches to machine translation have been used. For the massive translation of text, a statistical system has been trained and adapted to translate the text and transfer the semantic annotations into the target languages. The patent documents are semantically enriched and translated using the statistical system. The resulting multilingual documents are used to feed the databases of the retrieval system. On the other hand, a rule-based system is built in order to translate from (controlled) natural language to the semantic query language (SPARQL), in the interface.

The integration of different translation methodologies into the system has been crucial to increase its capabilities and make possible extended features and functionalities, with respect to preliminary version of the system. The preliminary version of the prototype, described in Deliverable 7.1 [MOL12c] had only original patent documents in the databases and the system was only available in English and French. A complete version of the prototype, described in Deliverable 7.2 [MOL12d], included resources also for German, and patent documents translated using the Statistical Machine Translation (SMT) system trained on the domain, and described in Deliverable 5.2 [MOL12a].

The news introduced with respect to previous versions of the prototype are:

1. A new process for statistical-based translation of patents that allows to transfer the semantic annotations and the original mark-up in the source documents to the target language (Section 2.1).
2. The updates on the retrieval architecture in order to improve the response time (Section 2.2).

3. A new querying approach for SPARQL generation based on the grammar – ontology interoperability automation (Section 2.3).
4. A new query grammar for the patents domain (Section 2.3).
5. The new functionalities integrated in the user interface in order to improve the usability of the application, such as the integration of the free-text search as a back-off mechanism for the query language (Section 2.2.1 and Section 3).

2 The Patents Case Study Prototype

The multilingual patents retrieval system (henceforth patents prototype) consists mainly of four modules (see Figure 1). First, the English sections of the original patent documents are semantically annotated using the pipeline especially designed for the ontologies on the domain. The annotated documents are preprocessed, cleaned and marked in order to translate the text using a statistical system trained on the biomedical domain. Both, original and translated texts are then merged into single file in order to feed the retrieval system's storage. Then, a specific query language has been designed for this use case in order to cope with the semantic annotation types and the relations in the domain ontologies.

The prototype is accessible on-line. It has a web-based graphical interface that allows for querying the system using a controlled natural language. The translation of the user queries, driven by the Grammatical Framework (GF), generates the SPARQL queries directly, as just a translation of the query to another target language. The results of the query are returned from the semantic repository and the document index, and they consist of both an RDF graph and the translated biomedical patent documents.

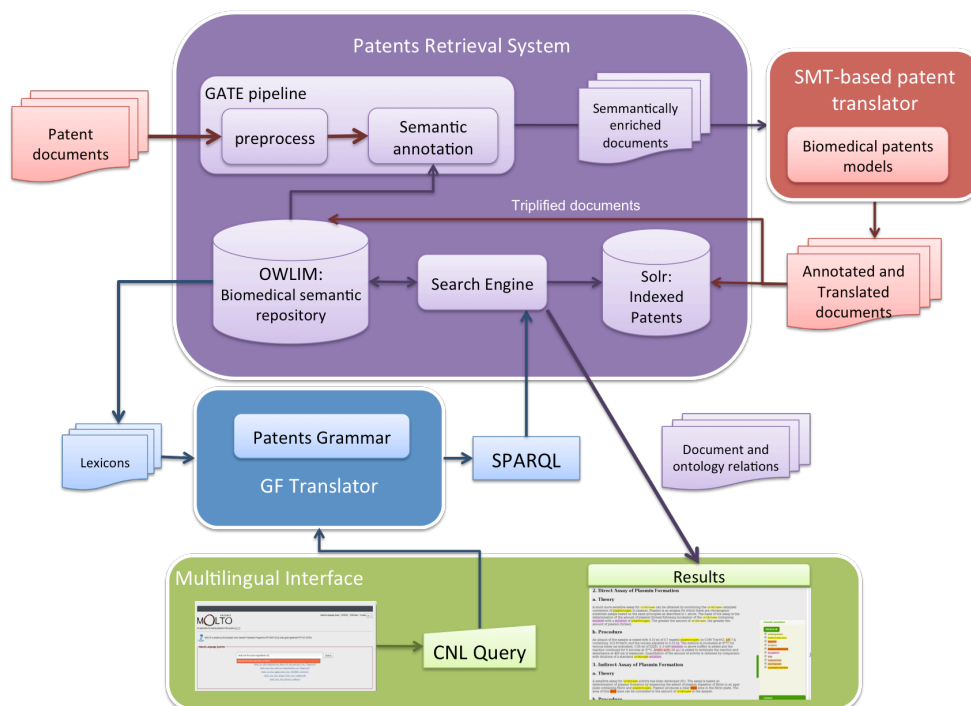


Figure 1: General architecture of the prototype

2.1 Patent corpus and translation

2.1.1 The biomedical patent datasets

The preparation of the patent corpus and the translation of the documents that feed the prototype databases is part of the work carried at UPC. For the patents case study we have used three different datasets.

The European Patent Office (EPO⁴) provided some parallel corpus containing the text of 66 patents belonging to the biomedical domain (IPC⁵ 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.

The statistical machine translation system requires a larger parallel dataset to train the models. For this purpose we used the MAREC corpus which contains not only European patents, but also documents from other international patent offices, dated from 1976 to 2008. Nonetheless, only the documents having text in English, French and German were used, and these were namely the EPO documents. The Deliverable 5.1 [MOL11] gives the description of this collection and the characteristics of the statistical machine translation (SMT) system built with these data.

The EPO provided also a website from where we downloaded 7,705 patent documents, also in the biomedical domain, them all dated from 2010 to 2012. This collection of documents constitutes the dataset used in the patent prototype.

The patent documents follow the normalized XML format defined by the EPO. This format consists of the following sections: the bibliographic data, an abstract, a description, a number of claims and references. The abstract, the description and the claims are always written in one of the three official languages, i.e., English (EN), German (DE) and French (FR), and sometimes they contain also the translation provided by the authors to any of the other two languages or both of them. We are using these documents to feed the patents retrieval system. To this end, the English sections of the patents are semantically annotated and automatically translated using the process described below, which translates semantically enriched text from the source language (EN) to the target (FR/DE).

From all the documents gathered from the EPO website, up to 4,485 out of the 7,705 documents contain at least one section with abstracts, claims or descriptions written in English. This is the final dataset that we annotated semantically according to concepts derived from several interrelated ontologies of the biomedical and patents domains. Table 1 gives a numerical description of the dataset, i.e., the number of documents having claims, description and abstract sections in English, German and French, respectively. As can be noted, the content of the documents can vary, some of them are bilingual or trilingual, not all of them have descriptions nor abstracts, and the number of claims varies as well.

The complete collection of files is available in the MOLTO svn repository⁶, and it consists of 1) the original patent documents, 2) the English sections of the patent doc-

⁴<http://www.epo.org/>

⁵<http://www.wipo.int/classifications/ipc/en/>

⁶<svn://molto-project.eu/patents-corpora/EPO-www-patents/>

uments semantically annotated, and 3) the automatic translations of claims, abstracts and descriptions into French and German merged into a single file with the original XML schema, defined by the EPO, and markup.

	Documents	Claims	Descriptions	Abstracts
English	4,485	62,638	3,832	2,518
German	2,047	32,007	192	80
French	2,011	31,487	130	44

Table 1: Number of sections in the patents prototype dataset

2.1.2 Translation of patent documents

The designed process for patents translation allows for building a translated document having the same XML structure as the original patent and having semantically enriched text. The purpose of this procedure is not only to translate the text, but also to preserve the semantic annotations during the translation process and keep the structure of the chemical compounds (we identified a total number of 1,097,243 compound instances in the documents, 243,823 different lexical entries), given that this is one of the key aspects of the biomedical texts.

The pipeline of the translation process is shown in Figure 2 and the example in Appendix A shows the transformations on the text at each step. The example shows first the original content of a patent document (1) that contains special sections such as an image and sub-indexes.

The first step in the pipeline consists of pre-processing this text in order to get rid of non-relevant marks, such as “” and “<i>”, because they harm the translation quality, but to preserve other structural markup such as enumeration and heads; encode the text in UTF-8 in order to preserve the symbols in the formulae and the compounds; and annotate the text with the semantic concepts. Example (2) shows the result of this process, which includes the UTF-8 symbols and the information about the *PARP* instance of the class *AnatomicalStructure*. Table 2 gives a summary of the semantic classes used and the number of instances found in the corpus; and Appendix F gives several lists of instances for these concepts.

In the second step, as shown in the diagram, the patent text is extracted from the sections in a structured manner. The resulting text is segmented and tokenized, as required by the translation system. This process is not trivial. On the one side, the characteristics of the domain require a special tokenization process. On the other side, sentences are in general too long and cannot be passed to the translator. The goal is to avoid the arbitrary segmentation of the sentences and improve the translation quality yet following predefined and prioritized clues in the text, such as enumerations and paragraph marks. In order to preserve the markup of the text, we make use of the “zone” and “wall” functionalities of the Moses translator [KSF⁺06]. This way, we can maintain the position of the marks in the

Concept	# Instances	Concept	# Instances
ActiveIngredient	285,192	AnatomicalStructure	2,170,330
Applicant	26,177	DiseaseOrDysfunction	1,218,063
DosageForm	241,279	Drug	160,698
Measurement	1,704,078	PharmaParam	39,330
Receptor	32,885	RouteOfAdministration	99,468

Table 2: List of semantic concepts and number of instances found in the dataset

text while certain degree of word reordering is still allowed during the decoding process. After this step, the structural marks are removed and the remaining text consists of the raw text shown in example (3).

During the third and fourth steps, the raw-marked text is translated using the SMT system and the translated text is post-processed in order to recover the original structure of the document (4), including original formatting, claims enumeration, document source and language metadata and semantic annotations. This process requires the original XML document as shown in the diagram.

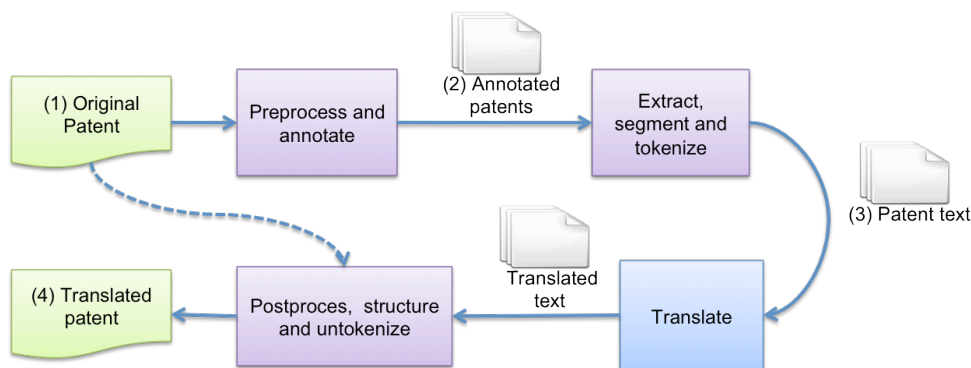


Figure 2: Patent document translation pipeline

The patent documents are translated using a variant of the SMT system described in Deliverable 5.2 [MOL12a], which was slightly modified in order to cope with the UTF-8 encoding used in this context. To do so, all the training datasets (*train* and *dev*) were converted into UTF-8 and the system was trained again. As a result, the evaluation results of the system, shown in Table 3 and Table 4, varies from the testset reported originally in D5.2. However, one might take into account that the results are not directly comparable to the ones reported in D5.2. First, the tokenization is different which is caused by the new codification and also the new functionalities developed in the evaluation tool to cope with the evaluation needs in MOLTO. And also, because the translations produced by Bing⁷

⁷<http://www.bing.com/translator>

and Google Translate⁸ are also different⁹.

It is worth to note also that the system was build originally to cope with patent claims and abstracts, and so that the parallel corpus described in D5.1. The language used in this

System	WER	PER	TER	BLEU	NIST	GTM-2	METEOR-st	ROUGE-S*	ULC
EN -> DE									
MOLTO-SMT	30.34	22.72	28.69	57.86	9.28	46.98	68.00	63.56	74.80
Bing	51.74	39.84	49.38	35.90	6.73	30.90	52.05	43.83	40.88
Google	31.79	22.85	29.79	62.53	9.77	49.58	70.11	68.39	77.61
EN -> FR									
MOLTO-SMT	24.96	17.07	23.51	64.70	10.00	49.45	77.19	73.45	74.92
Bing	39.18	29.07	36.45	46.75	8.49	35.76	61.88	59.09	47.42
Google	24.99	17.32	22.96	67.29	10.31	50.16	77.79	77.04	76.70

Table 3: Lexical evaluation of the UTF-8 - SMT system

System	CP-Oc(*)	CP-Op(*)	CP-STM-9	SP-Op(*)	SP-pNIST-5	ULC
EN -> DE						
MOLTO-SMT	54.57	60.54	35.99	60.54	7.26	97.86
Bing	39.22	44.10	24.70	44.10	5.88	71.90
Google	57.14	62.46	36.02	62.46	7.10	99.56
EN -> FR						
MOLTO-SMT	65.99	68.36	54.02	72.29	7.77	97.14
Bing	51.73	53.08	38.37	59.40	6.76	76.98
Google	67.78	69.64	55.81	74.88	8.02	1.00

Table 4: Syntactic evaluation of the UTF-8 - SMT system

2.1.3 The patent translation API

The source code for the whole pipeline is available at the MOLTO repository¹⁰. Processing and translating the whole bunch of documents takes several days if done sequentially. For this reason, the pipeline has been optimized to parallelize the processes when possible, depending on the characteristics of the computation environment and the number of documents. In order to facilitate its use, a main script runs the whole pipeline sequentially for a given set of input files.

On the one side, a simple web interface¹¹ has been set up to translate full patent documents and display the results. This demo uses only the SMT-based system and runs in a cluster of computation at UPC facilities.

⁸<http://translate.google.com/>

⁹Bing and Google update their systems periodically. These translations are dated April 2013.

¹⁰The source files can be found in <svn://molto-project.eu/patents-corpora/corpora-parser.tgz> along with the documents used in this prototype.

¹¹<http://nlp.lsi.upc.edu/molto> , {credentials molto/onlinepatenttranslation}

On the other side, a simple API has been set up in the MOLTO server to enable the remote use of the process, which can facilitate the integration with other tools. In contrast to the basic online translator, this service is more flexible and configurable. It is able to use both, the SMT and the hybrid systems (described in Deliverable 5.3[MOL12b]). It can translate not only a full patent document, but also a raw file and any raw text written by the user in, for instance, a text box in the interface. In order to speed up the translation process, the systems have been binarized and the phrase tables are filtered before the translation according to the user input. Furthermore, the systems have been updated in order to deal with multiple translation request enabling the possibility to use them from the online translator tool developed in WP2 and available in the GF cloud. The specification of this API is given in the Appendix B.

2.1.4 Syntactic measures for MT evaluation

The Asiya tool-kit [GM10a], developed at UPC, is an open-source framework for MT evaluation and meta-evaluation. It contains a rich set of evaluation metrics operating at different linguistic level and it uses several similarity measures.

The latest developments in Asiya include the possibility to provide already tokenized text, the integration of the MALT dependency parser [NHN⁺07], for which we obtained models for Catalan, Spanish, English and French, and the integration of the Berkeley Parser [PBTk06, PK07], for which we use a pre-trained model for German. These two new parsers have made possible the development of the syntactic evaluation measures, used for automatic MT evaluation in MOLTO, for French and German, in addition to English. In MOLTO, the Asiya tool-kit needed the development of specific functionalities to cope with the biomedical domain and the MOLTO requirements, such the tokenization of the text and the adjustment of the parsers for the measures described below.

Measures based on shallow parsing (SP) are available for all languages (English, French and German). The SP similarity is based on the lexical overlap according to the part-of-speech. For instance, $SP - O_p(NN)$ reflects the proportion of correctly translated singular nouns. The coarser measure $SP - O_p(*)$ computes the average lexical overlap over all parts of speech. We also use the NIST measure to compute accumulated and individual scores over sequences of ($n = 1..5$) parts-of-speech ($SP - NIST(i)_p - n$).

Measures based on constituency parsing (CP) are also available for all languages. They analyze similarities between constituent parse trees associated to automatic and reference translations. Constituent trees are obtained using the [CJ05] parser for English, the Bonsai v3.2 tool for French [CNDA10], and the Berkeley Parser for German. $CP - STM(i)_l$ measures calculate variants of the syntactic tree matching (STM) measure by Liu and Gildea (2005). For instance, $CP - STMi5$ calculates the proportion of length-5 matching sub-paths. $CP - O_p(t)$ computes the lexical overlap according to the part-of-speech ‘t’. And $CP - O_c(t)$ computes the lexical overlap according to the phrase constituent type ‘t’.

Measures based on dependency parsing (DPm) are available only for English and French. They capture similarities between dependency trees associated to automatic and reference translations. The pre-trained models for French were obtained from the French

Treebank [CCD10] and used in the Bonsai parser, which in turn uses the MALT parser (m). In contrast, the English text is processed with MALT parser [NHN⁺07] and MINIPAR [Lin98].

Finally, the ULC measure is a normalized arithmetic mean of metric scores. Its calculation implies the normalization of heterogeneous scores, some of them not even bounded, into in the range [0, 1]. As a consequence, the score constitutes a natural way of building a ranking rather than an overall estimation of the quality. This scheme has proven efficient when used with an appropriate combination of measures, as described in [GM10b].

2.2 Patents retrieval system, ontologies and semantic annotations

The patents retrieval prototype is an overlay of the MOLTO web-based retrieval system, that is based on the KRI¹² prototype [MOL10]. This system, developed and adapted by Ontotext, combines machine translation and multilingual retrieval of patents and has the potential to search through structural knowledge databases (the domain ontologies) and multilingual documents (the biomedical and multilingual patent documents).

The KRI is the data modeling and knowledge management core of the prototype. It is based on well-known platforms and tools, such as OWLIM¹³. The KRI allows also for building the conceptual models and knowledge bases, needed for developing the query subsystem, yet using the specialized datasets for the biomedical and patents domain, i.e., the patent classification taxonomies, the drugs lexicon, etc. Furthermore, the patents prototype integrates also Apache Solr¹⁴ as means for document snippeting and free text search (FTS).

This section describes the specific actions and new functionalities required to build the KRI retrieval system specialized for the patents use case: biomedical ontologies (Section 2.2.2), semantic annotations (Section 2.2.3) and patent query language (Section 2.3). Next, Section 3 details the user interface for document visualization.

Since the prototype version described in D7.2 [MOL12d], the patents retrieval system has been improved in a few directions. Figure 3 shows an updated view of the retrieval system (wrt. previous versions).

First, SPARQL generation was delegated completely to GF (Section 2.3). Then, free text search (Section 2.2.1) was added to the system, as a fall-back technology for queries that are not covered by the query language. This functionality uses Apache Solr, that is now also used for document snippeting and results in significantly higher speed response of the system (Section 2.2.5).

Finally, the domestic software products were updated to the last stable versions¹⁵. For

¹²<http://molto.ontotext.com/>

¹³OWLIM (<http://www.ontotext.com/owlim>) is a commercial RDF database management system, developed by Ontotext.

¹⁴<http://lucene.apache.org/solr/>

¹⁵with respect to the time of update - December, 2012

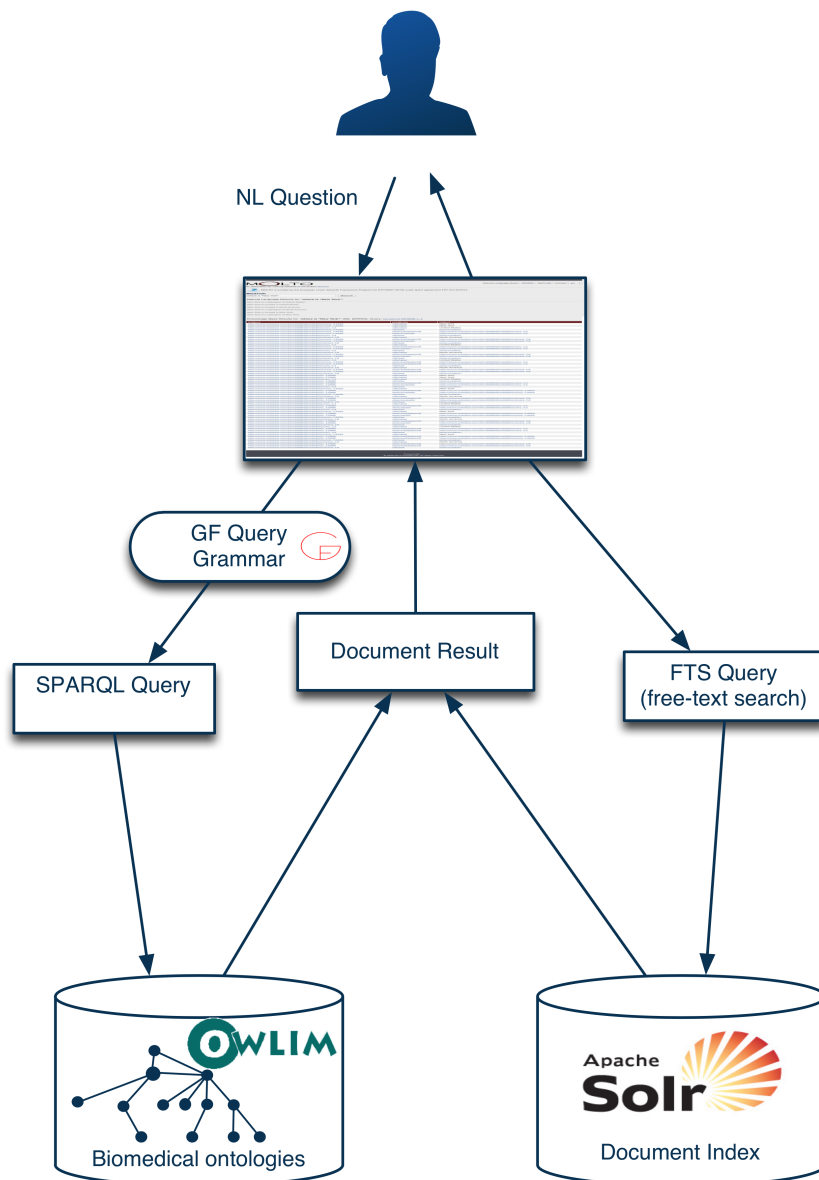


Figure 3: Patents retrieval system architecture with controlled natural language queries and free text search

instance, in the development of cultural heritage use case¹⁶ (WP8), we observed the need to use OWLIM 5 in order to make possible the use of federated queries to remote SPARQL end-points. Hence, all the retrieval systems in MOLTO (molto-kri, molto-patents and molto-cultural-heritage) were migrated to OWLIM 5 and Forest 1.4¹⁷. In the patents use case, this resulted in some small improvements in the user interface, including SPARQL queries syntax colouring. Federated SPARQL queries are also allowed, though not explored in this work package as in WP8.

2.2.1 Free text search

There has been a general observation that the controlled language lacks the extensive coverage of the human speech. Even in a closed domain, such as biomedical patents, a user may want to experiment with diverse queries and get frustrated with no reply to queries that could be considered straightforward for humans.

Fall-back mechanisms to the interpretation of natural language queries can vary. In the patents prototype, we have proposed full text search (FTS) as a feasible solution. The major motivation behind this decision is the fact that our system, apart from the knowledge from a semantic database, provides documents as results. However, this functionality is not a major focus of the work in WP7.

It is worth pointing out that the biomedical lexicons for the GF query grammars (e.g., drugs, active ingredients, etc.) were extracted from the ontologies. Therefore, not all existent drugs and other concepts from the patents were included in the query language. The use of the free text search functionality makes it possible to retrieve also the documents containing such entities.

Furthermore, the Solr index supports multilinguality. The English, French and German version of the documents are indexed separately. In the Solr terminology, each of the documents is indexed into three, very large and customized, text fields, one per language. In our case, these fields are “content_en”, “content_fr” and “content_de”, and the “id” is the actual patent number. The documents (which are initially trilingual), are split via the use of the XML metadata present in the “claims” tags, which was a design decision negotiated between UPC and Ontotext as part of the translation output format. Full text queries and queries for document snippets are executed only on the field of the index, corresponding to the language selected by the user. Hence, the resulting documents always contain the search phrase in the specific language, and so the snippets on the results page. An example of its use is given in Section 3.3.4 that shows the results obtained for the “Zofenopril” drug-name.

¹⁶<http://museum.ontotext.com/>

¹⁷Forest is a web-based framework for management of RDF datastores and semantically annotated documents, developed by Ontotext.

2.2.2 Ontologies for the biomedical domain

From the point of view of the semantic data, the patents prototype is based on the Exopatent¹⁸ project developed by Ontotext and Matrixware¹⁹. It combines several public ontologies and dictionaries to bring up a model for the biomedical patents domain. A full list of the loaded ontologies is given in D7.2 [MOL12d], Appendix A. The conceptual model for the final ontology is shown in Figure 4, where the data model contains the following ontologies: FDA Orange book²⁰, PROTON²¹, UMLS²², MeSH²³, SAR²⁴ and SAFEPat.

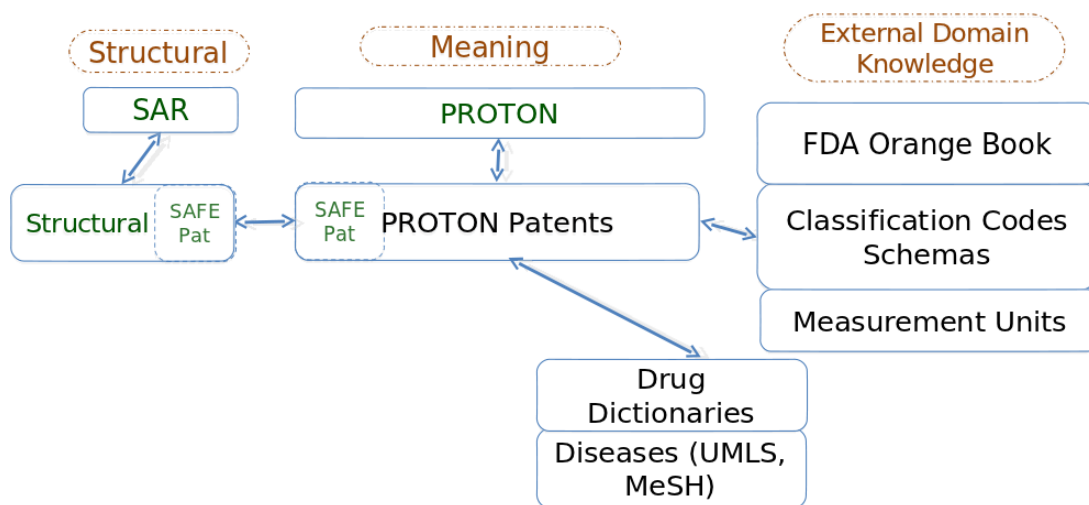


Figure 4: Conceptual model for the ontologies on the biomedical patents domain

It is important to stress that when we ask for e.g. the **active ingredients** of **ABSELECT**, we return all documents that are related to the active ingredients of this drug (in this specific case, to **AMPHOTERICIN B**). Intuitively one might expect to have both the drug and the active ingredient in one sentence in a returned document, but we do not keep relations such as “in the same sentence/paragraph” as in our system’s ontology. Rather, we look up for documents which were annotated to “mention” the subject of search, which we find to be **AMPHOTERICIN B**. Hence, the use of the biomedical ontologies, previously aligned for the purpose, allows us to find the relations between different entities, such as a drug and its active ingredients.

¹⁸<http://exopatent.ontotext.com/>

¹⁹Matrixware was a former participant in the MOLTO project.

²⁰<http://www.accessdata.fda.gov/scripts/cder/ob/default.cfm> and <http://www.accessdata.fda.gov/scripts/cder/ob/default.cfm>

²¹<http://www.ontotext.com/proton-ontology>

²²http://krono.act.uji.es/people/Ernesto/UMLS_SN_OWL

²³<http://www.nlm.nih.gov/mesh/>

²⁴The ontology of the Semantic Annotation Repository, Ontotext internal ontology for document presentation

2.2.3 Semantic annotations extracted from the patents metadata

In addition to the concepts listed in Table 2, that belong to the biomedical domain, we have reviewed the XML schema of the EPO patents²⁵ and decided to select and expose metadata that can be useful from the user's perspective, i.e., Patent Number, Application Number, Application Date and Publication Date, of a patent. As a result, we created four new semantic annotation types.

We extracted the entities from the documents, and added semantic annotations over them. Next, the new annotation instances were triplified (i.e., converted to RDF) so that they could be requested by SPARQL queries, and hence by natural language queries.

The example below shows the RDF interpretation of the new annotations on the patent number EP1326633B1. It has the application number "01966829", application date "20010914", publication date "20110316", and the patents's number (split in two parts) "1326633, B1".

This update allows a patent expert to query the system for a patent with specific patent number or application number, which we believe is a paramount functionality.

```
<http://molto-patents.ontotext.com/document/EP1326633B1> <http://proton.semanticweb.org/protonkm#mentions> <http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#PatentNumber.EP1326633B11> .
<http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#PatentNumber.EP1326633B11> <http://www.w3.org/2000/01/rdf-schema#label> "1326633" .
<http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#PatentNumber.EP1326633B11> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#PatentNumber> .
<http://molto-patents.ontotext.com/document/EP1326633B1> <http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#hasPatentNumber> <http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#PatentNumber.EP1326633B11> .
<http://molto-patents.ontotext.com/document/EP1326633B1> <http://proton.semanticweb.org/protonkm#mentions> <http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#PatentNumber.EP1326633B12> .
<http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#PatentNumber.EP1326633B12> <http://www.w3.org/2000/01/rdf-schema#label> "B1" .
<http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#PatentNumber.EP1326633B12> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#PatentNumber> .
<http://molto-patents.ontotext.com/document/EP1326633B1> <http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#hasPatentNumber> <http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#PatentNumber.EP1326633B12> .
<http://molto-patents.ontotext.com/document/EP1326633B1> <http://proton.semanticweb.org/protonkm#mentions> <http://proton.semanticweb.org/2006/05/patents#PublicationDate.EP1326633B1> .
<http://proton.semanticweb.org/2006/05/patents#PublicationDate.EP1326633B1> <http://www.w3.org/2000/01/rdf-schema#label> "20110316" .
<http://proton.semanticweb.org/2006/05/patents#PublicationDate.EP1326633B1> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://proton.semanticweb.org/2006/05/patents#PublicationDate> .
<http://molto-patents.ontotext.com/document/EP1326633B1> <http://proton.semanticweb.org/2006/05/patents#hasPublicationDate> <http://proton.semanticweb.org/2006/05/patents#PublicationDate.EP1326633B1> .
<http://molto-patents.ontotext.com/document/EP1326633B1> <http://proton.semanticweb.org/protonkm#mentions> <http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#ApplicationNumber.EP1326633B1> .
```

²⁵<http://docs.epoline.org/ebd/doc/ep-patent-document-v1-4.dtd>

```

<http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#ApplicationNumber.
EP1326633B1> <http://www.w3.org/2000/01/rdf-schema#label> "01966829" .
<http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#ApplicationNumber.
EP1326633B1> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://www.semanticweb.org/
ontologies/2008/7/Ontology1218740600570.owl#ApplicationNumber> .
<http://molto-patents.ontotext.com/document/EP1326633B1> <http://www.semanticweb.org/ontologies
/2008/7/Ontology1218740600570.owl#hasApplicationNumber> <http://www.semanticweb.org/
ontologies/2008/7/Ontology1218740600570.owl#ApplicationNumber.EP1326633B1> .
<http://molto-patents.ontotext.com/document/EP1326633B1> <http://proton.semanticweb.org/
protonkm#mentions> <http://proton.semanticweb.org/2006/05/patents#ApplicationDate.
EP1326633B1> .
<http://proton.semanticweb.org/2006/05/patents#ApplicationDate.EP1326633B1> <http://www.w3.org
/2000/01/rdf-schema#label> "20010914"

```

2.2.4 Semantic annotation over millions of patents

In the scope of MOLTO we experimented with semantic annotations over the whole MAREC patents corpus²⁶. It contains over 2.6 million patent documents pertaining to 1.3 million patents from the EPO with some content in English, German and French, and extended by documents from the WIPO²⁷. Some figures and descriptions of this corpus can be found in Deliverable 5.1[MOL11].

The idea behind this experiment was to obtain more featured data for training the SMT system for patents. Processing such unusual large dataset was challenging and it raised the need for a more scalable architecture. To this end, we designed a multithreaded document annotator tool for the GATE pipeline used in the patents domain. The number of threads is configurable and the state of the annotator's process was saved so that when the process crashed or failed, it was automatically resumed from or near the crash point²⁸.

2.2.5 Retrieval speed optimization

Before we started the final evaluation, we observed that the response time of the retrieval system was too high. After investigation, we found out that the creation of document snippets was the actual bottle neck. The reason was that in the previous version we created snippets for every single document through the GATE²⁹ library. The process was as follows: each retrieved patent was loaded in the memory, so that it was examined for a semantic annotation that was explicitly mentioned in the RDF result. Having the average size of 1.24 MB per document, and hundreds of annotations per document that had to be observed, the system's response time reached up to several seconds for the most complex queries. This timing is not acceptable for any commercial retrieval system, so we decided to switch to the Solr snippeting in order to keep the promise for a commercially viable prototype.

The Solr index is much faster, as the document content is indexed preliminary. Yet, we re-indexed the document content, so that each language of a document is observed

²⁶Data source: <http://www.ir-facility.org/prototypes/marec/statistics>

²⁷World Intellectual Property Organization, <http://www.wipo.int>

²⁸The final MAREC-annotated dataset is not available in the repository due obvious size problems.

²⁹<http://gate.ac.uk/family/embedded.html>

separately (the indexed texts were taken out of the translated texts using the Apache Tika³⁰).

Our Solr index stores only the patent text, without the semantic metadata. Therefore, the next challenge was to get the snippet by having only the document name and the instance of the annotation. An additional difficulty came from the fact that we deal with three languages, and the annotations have different RDF labels (i.e., texts) for each language.

As a solution, an auxiliary hash map index was created. For each annotation, in each document, the index stores a key following the pattern:

$$< document_id > ##### < annotation_instance > ##### < language >$$

and its value (i.e., the annotation text), as it appears in the patent text for the selected language. The index was used to retrieve the annotation text so that a snippet from the document can be requested directly from Solr by querying the annotation text. These steps improved the overall retrieval speed dramatically - about 5 times (exact speed measures weren't taken at the beginning). The exact measurements for the new methodology were performed on the updated system via Selenium tests and are discussed in Section 2.2.6.

2.2.6 Evaluation of the retrieval system

We have conducted a basic evaluation of the retrieval system in two directions: system response time and retrieval results for natural language queries.

System response times

We measured the system response times for NL queries executed on the prototype's user interface (UI), using Selenium³¹ and custom scripts. The default test browser was Firefox3; the experiment program was given 3G of memory.

We explored the following test scenario: click on all example queries shown in the home page of the patents prototype website (in English) and then, click on the "Search" button. The time in which the system returned results was measured. For all queries, the average time is presented below.

NL queries accuracy

We have conducted a large number of tests on the queries to verify whether the documents returned are correct responses to the natural language queries. Some of the results are presented on Table 7. The *precision* results are the highest possible, as expected for such a system. The reason for this is the lack of any variance in the possible answers, given the control we have over the SPARQL queries. As far as it concerns *recall*, the documents we return are guaranteed to "mention" a specific object from the ontology. Usually we have its label and define our search query via it. Then, we return all documents that were

³⁰<http://tika.apache.org/>

³¹<http://docs.seleniumhq.org/http://docs.seleniumhq.org/>

Attempt	Opening of Home Page (ms.)	Avg. Response Time to NL Query (ms.)
Test 1	572.36	72.29
Test 2	444.00	76.64
Test 3	381.86	77.00
Test 4	452.71	74.50
Test 5	472.43	73.36

Table 5: Average system response measures in ms.

Query	Avg. response time (ms.)
what is the expiration date of the patent for PACERON	74.60
what is the application number for EP2024375B1	73.60
all the patents that mention GABITRIL	77.60
what are the active ingredients of BACLOFEN	69.00
give me all the drugs with the market RX	75.20
all the drugs with the active ingredient ACARBOSE	73.00
all information about AMPICILLIN	84.60
what is the patent with the application number 07252109	72.00
show me the drugs with the therapeutic equivalence code AA	85.20
what is the application date of EP1771422B1	72.20
what is the strength of SELENIUM SULFIDE	72.83
what is the approval date of the patent for REBETOL	71.20
patents that mention SELENIUM SULFIDE and AMPICILLIN	75.20

Table 6: Average system response measures in ms.

annotated to contain the specific object. Any possible errors would be due to incorrect entity recognition during the semantic annotation phase, which is not broadly explored in the scope of MOLTO.

Next, the updated query language maps the relations in the ontology much better than in the previous version of the prototype.

In the above table “all OK” stands to denote applicable results(rdf facts and documents) and “NA” denotes if the query is not supposed to return documents.

2.3 GF grammars for generation of SPARQL

For the final version of the patent prototype, we applied for the first time the novel approach of SPARQL generation by GF grammars. Namely, this is a GF translation model able to translate natural language queries into SPARQL, considering SPARQL as yet another “natural” language with its own concrete grammar (see Figure 5).

In gross, instead of a grammatical representation of a sentence using the Resource Grammar Library[Ran09], the new approach provides a SPARQL representation. Hence, the translation of natural language to SPARQL is fully done by the GF engine. This

Query	Result	Documents returned
patents that mention SELENIUM SULFIDE and AMPICILLIN	all apply	http://molto-patents.ontotext.com/document/EP2384753A1 http://molto-patents.ontotext.com/document/EP1469876B1 http://molto-patents.ontotext.com/document/EP2068873B1
what are the patents that mention PHENAZINE and URETHRAL	all apply	http://molto-patents.ontotext.com/document/EP2298767A1 http://molto-patents.ontotext.com/document/EP2089369B1
what is the application date of EP1771422B1	OK	http://molto-patents.ontotext.com/document/EP1771422B1
what is the strength of SELENIUM SULFIDE	all apply	http://molto-patents.ontotext.com/document/EP2384753A1 http://molto-patents.ontotext.com/document/EP1469876B1 http://molto-patents.ontotext.com/document/EP2343304A1 http://molto-patents.ontotext.com/document/EP2286820A1 http://molto-patents.ontotext.com/document/EP2068873B1
drugs with the active ingredient ACARBOSE	all OK	50 Documents
show me all the routes of administration of THIAMINE HYDROCHLORIDE	all OK	NA
what are the dosage forms of A-HYDROCORT	all OK	NA

Table 7: Correctness of results to NL queries

automation saves lots of additional efforts with respect to providing a mapping from natural language to SPARQL.

In the previous version of the prototype, we used a domain specific language, designed by Ontotext for this purpose. A mapping-rules example of the previous version of the grammar is available at the MOLTO repository [svn://molto-project.eu/wp7/molto-patents/natural-language-queries/resources/mapping-rules](http://molto-project.eu/wp7/molto-patents/natural-language-queries/resources/mapping-rules).

It was obvious that these rules would require human maintenance and updates in order to extend the coverage of the language or cope with any update of the ontologies, as well as the parser for the rules itself. In contrast, the new GF-based SPARQL-grammars approach minimizes the need for this maintenance.

The core grammar for the new presentation is the YAQL (“yet another query language”)³², which is developed by UGOT as an extension of WP4. The query grammar for the patents case study is located at [svn://molto-project.eu/wp7/query/grammars](http://molto-project.eu/wp7/query/grammars).

In the Cultural Heritage case study (WP8), the SPARQL generation by GF evolves in a functional representation, which brings the automation to a higher level. Details can be found in D8.3[MOL13].

From the back-end perspective, the system runs a GF process and the SPARQL is generated by a single translation command. The following example

```
PatentQuery> p -lang=PatentQueryEng -cat=Query
"show me the patents that mention AMPICILLIN" | 1 -lang=PatentQuerySPARQL
```

returns the SPARQL query shown below (the \$n symbol is used to be an indicator of

³²[svn://molto-project.eu/wp4/YAQL](http://molto-project.eu/wp4/YAQL)

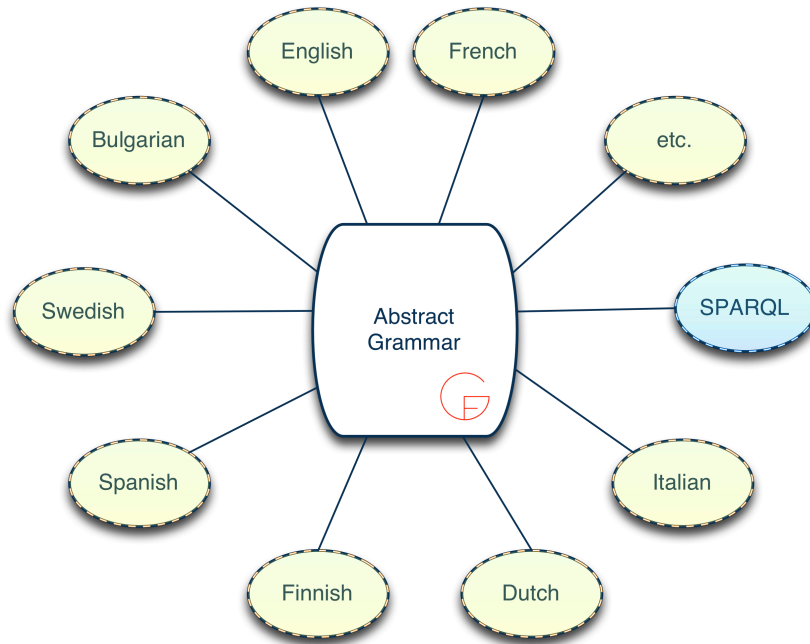


Figure 5: SPARQL as yet another GF language

a new line, which we process in the back-end in order to return a well formed SPARQL query).

```

PREFIX pkm: <http://proton.semanticweb.org/protonkm#> $n
PREFIX psys: <http://proton.semanticweb.org/protonsys#> $n
CONSTRUCT { ?doc <http://proton.semanticweb.org/protonm#mentions> ?thing . } $n
WHERE { $n
?thing psys:mainLabel " AMPICILLIN " . $n
?doc pkm:mentions ?thing . $n
} ;
  
```

2.3.1 The patents query language

Compared to the query language described in D7.2, the current one is improved in terms of coverage and compliance to the patent domain ontology that is behind the information retrieval system. For these reasons, the layered representation that builds on the Query Library, was abandoned in favor of a stand-alone implementation that is faithful to the requirements. In this way, the current grammar is an extension to the grammar described in D7.1, in order to cope with the latest changes in the ontology. In addition, the current grammar provides a concrete syntax corresponding to SPARQL.

Since the grammar is adapted to the domain, the constructors from the abstract syntax describe individual queries. So, the SPARQL mappings are easy to make, since they are

written in a gap-filling fashion, by specifying the query with spaces for the arguments. For example, the operation that builds the queries that ask about patents mentioning a certain concept is as follows:

```
oper mentionP : Str -> {s : Str} = \s1 ->
  {s = "PREFIX pkm: <http://proton.semanticweb.org/protonkm#> $n
    PREFIX psys: <http://proton.semanticweb.org/protonsys#> $n
    CONSTRUCT { ?doc
<http://proton.semanticweb.org/protonkm#mentions> ?thing . } $n
    WHERE { $n
      ?thing psys:mainLabel "++ s1 ++". $n
      ?doc pkm:mentions ?thing . $n} ;"} ;
```

In addition to the direct generation of SPARQL from the patent query grammar, the natural language generation has been optimized for French, rendering more fluent queries. For example, for "what is the publication date of EP2033641B1", the French query grammar now returns "quelle est la date de publication d' EP2033641B1" instead of "qu' est la date de publication d' EP2033641B1".

All in all, the current patent query grammar improves over the previous ones (from D7.1 and D7.2) in terms of coverage, quality of the natural language generated and by providing a direct mapping of the natural language queries to SPARQL, which favors scalability and domain adaptation. In addition to this, a large dictionary of terms has been mapped to grammar lexicon and in this way, the grammar provides all the necessary support for the query part of the information retrieval system.

2.3.2 Better coverage of the ontology by GF

In D7.2, we pointed that the natural language defined by the query grammar suffered from ambiguities and inconsistencies. We reduced these to a minimum, and also extended the grammar with several new and more precise natural language queries.

In previous versions of the patents, the lexicons used in the grammar were populated from the ontologies, that were available only in English. These lexicons did not cover the totality of the query-able content of the patents database, which limited the usability of the prototype.

The final grammar includes eight new lexicons (i.e., the dictionaries shown in Table 8), which allowed also for larger flexibility on the query language itself. Each lexicon corresponds to that semantic annotation types whose instances were of interest. The lexical entries were extracted from the annotated documents described in Section 2.1.1, which were originally in English.

In addition, the translation process described in Section 2.1.2 allowed us to obtain annotated items also in French and German, and consequently, we have been able to build also the lexicons for these languages, which were not available before. The benefit is notable, since the translation process allows us to build non-English query grammars entirely in the user's language, without the need to use English terms for the named entities.

Annotation (Entity) type	Examples	Dictionary size
DRUG	FLUONID, HEXADROL	5,529 entries
ACTIVE_INGREDIENT	BUDESONIDE, MORPHINE SULFATE	1,803 entries
ROUTE_OF_ADMINISTRATION	DENTAL, CARDIAC	53 entries
APPLICANT	ASTRAZENECA, BAYER PHARMS	1,822 entries
APPLICATION_NUMBER	10184202, 09814154	4,609 entries
PATENT_NUMBER	EP2219477B1, EP2382981A3	4,609 entries
TE_CODE	AA, AB, BX	17 entries
MARKET	DISCN, OTC, RX	3 entries

Table 8: Dictionary types for the data GF grammars

The extension of the query language introduced new useful queries. For example, we added the following queries:

```

what are the drugs with the active ingredient ACTIVE_INGREDIENT
what are the drugs with the therapeutic equivalence code TE_CODE
what is the application date of PATENT
what is the application number for PATENT
what are the patents that mention X
what are the patents that mention Y and Z

```

where X, Y and Z are one of the above types {DRUG, ACTIVE_INGREDIENT, APPLICANT, ROUTE_OF_ADMINISTRATION}.

A selected list of lexical items for the ontology classes can be seen in Appendix F. The complete list of concept instances (i.e., drugs, active ingredients, patent numbers, etc.) is available at svn://molto-project.eu/wp7/resources/lexicons.

Our observation is that the more the grammar approximates the RDF-predicate relation's names, the easier and the more precise SPARQL queries can be introduced.

3 The On-line User Interface for the Patents Retrieval Prototype

The patents prototype can be accessed on-line at:

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

The general workflow of the application is shown in Figure 3 (Section 2.2). We combined a number of diverse technologies to come up with a multilingual search engine for semantically annotated patents.

The interface allows for querying the system in the EPO official languages, i.e., English, German and French. The queries are written using the controlled natural language defined by the GF query grammar for patents described in the previous sections. The GF engine is used to parse the user's query to abstract syntax and to translate it to SPARQL syntax (as described in Section 2.3). Then, the SPARQL query is executed against the semantic repository. As a result, the user receives a subset of documents with snippets and RDF facts. Finally, the user interface displays a browsable list of domain concepts and documents that allows for navigating the ontology and inspecting the patent documents.

For those cases in which the user's query cannot be parsed by GF, we have enabled the FTS (described in Section 2.2.1) that uses Solr's indexing and returns to the user only document snippets that contain the keywords in search.

Regarding the graphical user interface, some deficiencies were detected in the previous version of the prototype, these being the main ones:

- inaccurate possible queries,
- insufficient coverage of the ontology relations by the query language,
- document snippets in English only,
- semantic annotations in English only, and
- slow response time of the application.

Our goal for the final prototype was to overcome these issues and come up with an online application that is commercially viable. We believe we have made significant improvements on the points listed above. The results are discussed in the next sections.

3.1 GUI updates

Among the updates towards the improvement of the user experience, we count the perceivable speed-up due to Solr snippeting and a couple of other minor GUI speed-ups. For details, see Section 2.2. Furthermore, we made the following improvements to the interface.

3.1.1 Multilingual annotations

Unlike the previous version of the prototype, in the final version the semantic annotations are transmitted to the translated sections of the patent documents. The process that lead to this advancement is described in Section 2.1. From semantic data’s perspective, this is a huge step forward. Figure 6 shows an excerpt of French text that has been automatically annotated.

This step sets up the grounding for further research on how well the annotations are transferred and whether machine translation could be used to enrich semantic annotations with labels in different languages.

Les composés de l'invention peut être utilisé en combinaison avec les agents pour améliorer de la nicotine privation et réduire de la nicotine manque : i) de la nicotine d'une thérapie de substitution par exemple un sublinguale dans la formulation de de la nicotine de bêta-cyclodextrine et de la nicotine patches; et ii)

Les composés de l'invention peut être utilisé en combinaison avec les agents pour améliorer alcool de sevrage et réduire alcool manque : i) récepteur de NMDA antagonistes par exemple acamprosate ; ii) récepteur GABA agonistes par exemple tetrabamate ; et iii) Opioid récepteur antagonistes par exemple la naltrexone .

Les composés de l'invention peut être utilisé en combinaison avec les agents afin d'améliorer la substance opiacée privation et réduire opiacé manque : i) opioïde mu récepteur agoniste opioïde / récepteur kappa antagoniste par exemple la buprénorphine ; ii) récepteur opioïde antagonistes par exemple la naltrexone ; et iii) vasodilatatrice au RECEPTOR seurs par exemple la lofexidine.

Les composés de l'invention peut être utilisé en combinaison avec les agents destiné au traitement ou à la prévention du sommeil troubles : i) les benzodiazépines par exemple le témazépam , lormetazepam , estazolam et triazolam ; ii) non-benzodiazépine somnifères par exemple zolpidem, zopiclone, zaleplon et indiplon ; iii) les barbituriques par exemple aprobarbital, le butobarbital , le pentobarbital , secobarbital et phenobarbital; iv) antidépresseurs ; v) d'autres sédatifs-hypnotiques par exemple chloral hydrate et chlorméthiazole.

Figure 6: Annotations transmitted to French

3.1.2 Document source exposed

Another minor improvement is the exposure of the "Document Source" of a patent text. This makes clear the origin of the text *in the chosen language*, i.e., original or translated by the automatic systems. Currently, the two possible options are "EPO" vs. "MOLTO-SMTv1", as indicated in Figure 7.

3.2 NL query examples

The controlled language of the prototype was extended and improved. As a result, we have new queries that cover the ontology much better (the most common ontology relations are covered by the GF grammars). Some of the new queries are essential for the usability of the prototype. For example, it is now possible to search a patent by its number, or by its application number. We also introduced series of queries like *give me all the drugs with the market RX*, where we collect drugs with common characteristics (e.g. market, active ingredient, therapeutic equivalence (TE) code, dosage form, etc).

Table 9 presents several query examples taken from the UI that the user can use to query the system. As in previous versions, the interface has an auto-complete functionality that can predict and suggest the applicable lexical types and entries (e.g., DRUG, TE_CODE,

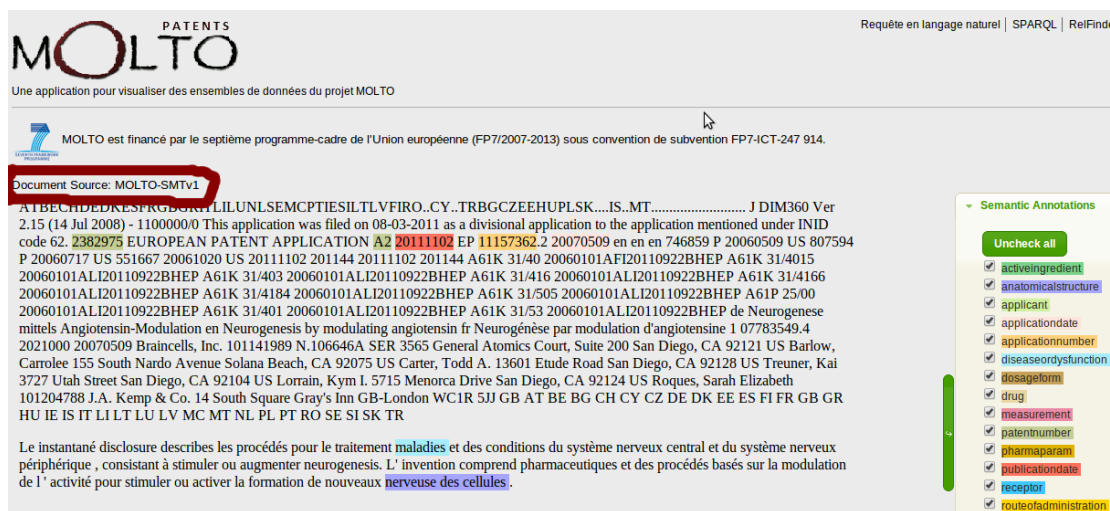


Figure 7: Patent document source

and so on) for each possible query. The table shows also the number of results that are currently returned (*NA* stands for “inapplicable result type for this particular query”).

Query	# of documents returned	# of RDF triples
all the patents that mention GABITRIL	6	6
what is the expiration date of the patent for PACERONE	NA	4
what is the application number for EP2024375B1	NA	1
what are the active ingredients of BACLOFEN	*43	2,262
give me all the drugs with the market RX	*12	1,421,934
all the drugs with the active ingredient ACARBOSE	*50	1,905
all information about AMPICILLIN	*91	308
what is the patent with the application number 07252109	1	1
show me the drugs with the therapeutic equivalence code AA	*86	769,172
what is the application date of EP1771422B1	NA	1
patents that mention SELENIUM SULFIDE and AMPICILLIN	3	12

* out of the top 100 RDF facts

Table 9: Query examples with results counts

3.3 User Scenarios

In this subsection we list some interesting user scenarios that are new to the system (or were not mentioned previously).

3.3.1 Query types

Among the wide variety of new queries added to the query language, the most powerful one is the *and*-query, meaning a query of the type:

show me all patents that mention X and Y

where X and Y have one of the types: Drug, Applicant, ActiveIngredient, and RouteOfAdministration (see Table 8 for details on the number of entities of each of the four types, and Appendix F for a list of instances).

The *and*-query allows a user to search for two different entities of the types mentioned above, which offers to the user the possibility to build up to 42 million different queries.

Figure 8 shows an example of an *and*-query and the results returned: the RDF triples and three documents. We also demonstrate that the two semantically annotated instances are located in the documents. Figure 9 shows the occurrences in patent number EP2068873B1.

Search

Knowledge Base Results for "patents that mention SELENIUM SULFIDE and AMPICILLIN" (12) (SPARQL Query: PREFIX pkm:...)

subject	predicate	object
http://molto-patents.ontotext.com/document/EP2384753A1	pkm:mentions	http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#ActiveIngredient.SELENIUM_SULFIDE_T.1528
http://molto-patents.ontotext.com/document/EP2384753A1	pkm:mentions	http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#FDA_DrugName.AMPICILLIN_T.513
http://molto-patents.ontotext.com/document/EP1469876B1	pkm:mentions	http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#ActiveIngredient.SELENIUM_SULFIDE_T.1528
http://molto-patents.ontotext.com/document/EP1469876B1	pkm:mentions	http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#FDA_DrugName.AMPICILLIN_T.513
http://molto-patents.ontotext.com/document/EP2068873B1	pkm:mentions	http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#ActiveIngredient.SELENIUM_SULFIDE_T.1528
http://molto-patents.ontotext.com/document/EP2068873B1	pkm:mentions	http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#FDA_DrugName.AMPICILLIN_T.513
http://molto-patents.ontotext.com/document/EP2384753A1	pkm:mentions	http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#FDA_DrugName.SELENIUM_SULFIDE_T.4819
http://molto-patents.ontotext.com/document/EP2384753A1	pkm:mentions	http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#FDA_DrugName.AMPICILLIN_T.513
http://molto-patents.ontotext.com/document/EP1469876B1	pkm:mentions	http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#FDA_DrugName.SELENIUM_SULFIDE_T.4819
http://molto-patents.ontotext.com/document/EP1469876B1	pkm:mentions	http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#FDA_DrugName.AMPICILLIN_T.513
http://molto-patents.ontotext.com/document/EP2068873B1	pkm:mentions	http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#FDA_DrugName.SELENIUM_SULFIDE_T.4819
http://molto-patents.ontotext.com/document/EP2068873B1	pkm:mentions	http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#FDA_DrugName.AMPICILLIN_T.513

Documents (3)

EP2384753A1

Some exemplary anti-fungal agents include imidazoles, FK 463, amphotericin B, BAY 36-9502, MK 991, pradimicin, UK 292, butenafine, chitinase and 501 cream, Acrisorcin; Ambruticin; Amorolfine, Amphotericin B; Azaconazole; Azaserine; Basifungin; Bifonazole; Biphenamine Hydrochloride; Bispirithione Magsulfex; Butoconazole Nitrate; Calcium Undecylenate; Candicidin; Carbol-Fuchsin; Chloridantoin; Clotriprox; Clotriprox Olamine; Clotriprox; Clotrimazole; Clotrimazole; Cuprimyxin; Denofungin; Dipyrithione; Doconazole; Econazole; Econazole Nitrate; Enilconazole; Ethonam Nitrate; Fenticonazole Nitrate; Filipin; Fluconazole; Flucytosine; Fungimycin; Griseofulvin; Hamycin; Isoconazole; Itraconazole; Kalafungin; Ketoconazole; Lomofungin; Lydimycin; Mepartricin; Miconazole; Miconazole Nitrate; Monensin; Monensin Sodium; Nafifine Hydrochloride; Neomycin Undecylenate; Nitafate; Nifumrone; Nitraamine Hydrochloride; Nystatin; Octanoic Acid; Orconazole Nitrate; Oxiconazole Nitrate; Oxifungin Hydrochloride; Parconazole Hydrochloride; Partricin; Potassium Iodide; Proclonol; Pyrrithione Zinc; Pyrrolinrin; Rutamycin; Sanguinarium Chloride; Saperconazole; Scopafungin; Selenium Sulfide; Sinefungin; Sulconazole Nitrate; Terbinafine; Terconazole; Thiram; Ticlatone; Tioconazole; Tolcilate; Tolindate; Tolnaffate; Triacetin; Triafungin; Undecylenic Acid; Viridofulvin; Zinc Undecylenate; and Zinoconazole Hydrochloride.

EP1469876B1

Exemplary anti-fungal agents include, but are not limited to, terbinafine hydrochloride, nystatin, amphotericin B, griseofulvin, ketoconazole, miconazole nitrate, flucytosine, fluconazole, itraconazole, clotrimazole, benzoic acid, salicylic acid, and selenium sulfide. The antimicrobial cationic peptide may also be used in combination with anti-viral agents. Exemplary anti-viral agents include, but are not limited to, amantadine hydrochloride, rimantadin, acyclovir, famciclovir, foscarnet, ganciclovir sodium, idoxuridine, ribavirin, sorbivudine, trifuridine, valacyclovir, vidarabin, didanosine, stavudine, zalcitabine, zidovudine, interferon alpha, and edoxudine.

EP2068873B1

Examples of antifungal agents suitable for use in the present invention include, but are not limited to, azoles, such as miconazole, clotrimazole, ketoconazole, oxiconazole, econazole, sulconazole, fluconazole, and itraconazole; allylamines, such as naftifine and terbinafine; benzylamines, such as butenafine; polyenes, such as nystatin and amphotericin B; thiocarbonates, such as tolinaffate; sulfides, such as selenium sulfide; nitrogen-containing heterocycles, such as clotriprox, and combinations of two or more antifungal agents. In another preferred embodiment of the invention, the pharmaceutical composition further comprises a combination of one or more antibacterial agent and one or more antifungal agents.

Figure 8: Results of the *and*-query: patents that mention SELENIUM SULFIDE and AMPICILLIN

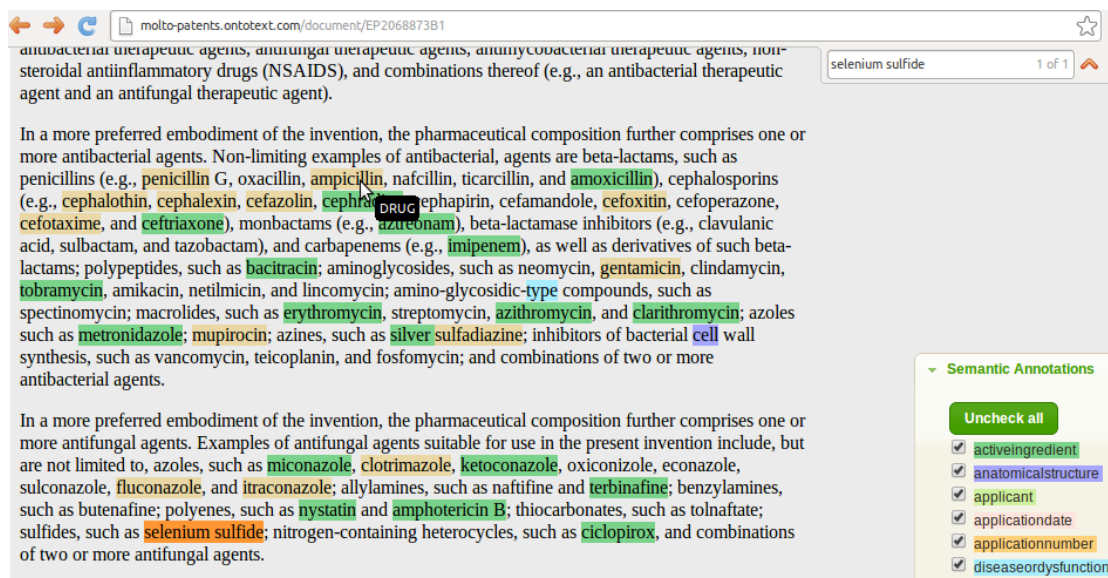


Figure 9: EP2068873B1 - a patent document retrieved for the query `patents that mention SELENIUM SULFIDE and AMPICILLIN`

3.3.2 Semantic data exploration

The semantic annotations³³ are metadata over an existing excerpt of a text document. Each semantic annotation has these two properties: a class (from the top-level of the ontology) to which it belongs to, and a unique URI that represents it (i.e., an instance). Other metadata is possible but optional.

Since the patents prototype is based on the KRI project³⁴, it facilitates the exploration of semantic data. We have provided the ability to follow the related semantic data about the instance by clicking on the annotation’s highlights (i.e., a browsable semantic graph). The semantic data gives richer and valuable information to the user. Let us explore the following use case to show it.

A user searches for the patents that mention the drug `ONDANSETRON` and opens a document that contains the drug, e.g. EP2086643B1. The user selects *drugs* from the “Semantic Annotations” menu (the check list menu bar in the right of the web page). The “Check all” radio button shows all the semantic annotation types.

Next, the user searches for the specific occurrence of `ONDANSETRON` (in the current version of the prototype one should use the browser’s search to find the exact occurrence of a string in the patent).

Once done, the user will be navigated to a paragraph that discusses about treatment and prevention of `bulimia`. The disease is semantically annotated. When the user selects *diseaseordysfunction*, the mention of `bulimia` is highlighted (see Figure 10). The user can then click on the annotation’s highlight and get redirected to a new page that shows

³³<http://www.ontotext.com/kim/semantic-annotation>

³⁴<http://molto.ontotext.com>

all occurrences of the bulimia’s instance. It is the *subject* in triples from the ontology loaded on the prototype.

Bulimia RDF Rank RDF Search and Explore

Eating an excess amount of food in a short period of time, as seen in the disorder of BULIMIA NERVOSA. It is caused by an abnormal craving hunger also known as "ox hunger".

Source: <http://linkedlifedata.com/resource/umls/id/C0006370>

Subject (12) Predicate Object All Download in: [JSON](#) | [RDF](#)

Statements in which the resource exists as a subject. Named Graph: [All](#) Language: [English](#) Inference: [Ex](#)

Predicate	Object
rdf:type	http://linkedlifedata.com/resource/semanticnetwork/id/T046
rdfs:comment	Eating an excess amount of food in a short period of time, as seen in the disorder of BULIMIA caused by an abnormal craving for food, or insatiable hunger also known as "ox hunger".
rdfs:label	BULIMIA Binge Eating Bulimia Bulimia [Disease/Finding] Bulimias Eating, Binge bulimia
psys:mainLabel	Bulimia
psys:generatedBy	http://ontotext.com/UMLSDump
skos:broader	http://linkedlifedata.com/resource/umls/id/C0004936

Figure 10: bulimia as *subject* in the RDF triples

Another interesting view is at the Object tab. It shows the RDF facts in which the same instance is an *object*. This usually gives the list of documents that “mention”³⁵ the entity. In the specific case we get the list of documents that have bulimia annotated in them, as in Figure 11. Hence, users are able to explore lots of linked information related to their search.

3.3.3 Single word query

It was observed that a novel user of the system may get frustrated because the lack of single-word queries, e.g., a drug name or a patent number. Hence, we added single-word queries in the GF query grammar, so that a user can directly ask for a patent number, a drug, an active ingredient, etc. An example is shown in Figure 12. The input box integrates also a predictive text mechanism for single substances and patent numbers. The latter saves both time and the possibility of spelling mistakes, which otherwise would have brought the user to free text search (see Section 3.3.4).

³⁵<http://proton.semanticweb.org/protonm#mentions>

Bulimia RDF Rank

Eating an excess amount of food in a short period of time, as seen in the disorder of BULIMIA NERVOSA. It is caused by an abnormal hunger also known as "ox hunger".

Source: <http://linkedlifedata.com/resource/umls/id/C0006370>

Download in: [JSON](#)

Statements in which the resource exists as a object. Named Graph Language Info

Subject	Predicate
http://molto-patents.ontotext.com/document/EP2283841A1	pkm:mentions
http://molto-patents.ontotext.com/document/EP1809597B1	
http://molto-patents.ontotext.com/document/EP2205560B1	
http://molto-patents.ontotext.com/document/EP2295066A1	
http://molto-patents.ontotext.com/document/EP1481969B1	
http://molto-patents.ontotext.com/document/EP2272847A1	
http://molto-patents.ontotext.com/document/EP1797088B1	
http://molto-patents.ontotext.com/document/EP1942106B1	
http://molto-patents.ontotext.com/document/EP2086643B1	

Figure 11: bulimia as *object* in the RDF triples

Unfortunately, it is impossible to have in the ontology all the drugs that are mentioned in the patent documents. There will always be a chance of a new drug in a so far unseen document. In the system described so far, it was not possible to query for entries that were not in the data lexicons extracted from the domain ontology. As a workaround to this issue we added free text search functionality to the system.

3.3.4 FTS queries

This version of the prototype supports free text search in the three languages (English, French and German). It serves as a fall-back mechanism for queries that the GF engine cannot parse and also for drugs names and chemical substances, compounds, etc., that are not in the ontology (and hence neither in the GF query grammar).

Figure 13 shows a free text search for the medicine **Zofenopril** that occurs in patents, but is not part of the ontology. The system returned 37 documents with sentence or paragraph snippets in the language of the user. The snippets explore the occurrences of the search phrase in the document.

Natural Language Queries

Famotidine

FAMOTIDINE PRESERVATIVE FREE (PHARMACY BULK)
FAMOTIDINE PRESERVATIVE FREE IN PLASTIC CONTAINER
FAMOTIDINE PRESERVATIVE FREE
FAMOTIDINE
FAMOTIDINE

[what is the approval date of the patent for NEDROL](#)
[what is the expiration date of the patent for PACERONE](#)
[what is the strength of SELENIUM SULFIDE](#)
[what is the patent with the application number 07252109](#)
[what is the application number for EP2024375B1](#)
[what is the application date of EP1771422B1](#)
[give me all the drugs with the market RX](#)
[all the drugs with the active ingredient ACARBOSE](#)
[show me the drugs with the therapeutic equivalence code AA](#)

Figure 12: Single drug search (by GF)

molto-patents.ontotext.com/search?q=Zofenopril&_form=%2Fsearch

MOLTO PATENTS Natural Language Query | SPARQL | RelFin

An application for viewing datasets of the project [MOLTO](#)

MOLTO is funded by the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement FP7-ICT-247914.

Search

Zofenopril

Knowledge Base Results for "Zofenopril" (0)

Documents (37)

[EP1844028B1](#)
No. 4,316,906 with [zofenopril](#) being preferred. Other examples of mercapto containing ACE inhibitors that may be employed, herein include rentin (Santen) disclosed in Clin. Exp. Pharmacol. Physiol. 10:131 (1983) ; as well as pivopril and YS980.

[EP2377532A1](#)
In one preferred embodiment of the present invention, the ACE inhibitor binds to the zinc-binding ligand of the active site of ACE via a sulphydryl captopril, [zofenopril](#) and/or alacepril, or a pharmaceutically acceptable salt thereof. In another preferred embodiment of the present invention, said to the zinc-binding ligand of the active site of ACE via a phosphinyl group, such as fosinopril or a pharmaceutically acceptable salt thereof. In one embodiment of the present invention, the ACE inhibitor binds to the zinc-binding ligand of the active site of ACE via a sulphydryl group, such as and/or alacepril, or a pharmaceutically acceptable salt thereof. In another preferred embodiment of the present invention, said ACE inhibitor binds ligand of the active site of ACE via a phosphinyl group, such as fosinopril or a pharmaceutically acceptable salt thereof.

[EP2301936A1](#)
The remedy for hypertension include, for example, 1) thiazide diuretics such as chlorothalidon, chlorothiazide, dichlorofenamide, hydrofluorothiaz hydrochlorothiazide et al; loop diuretics such as bumetanide, ethacrynic acid, flosemide, tolusemide et al; sodium diuretics such as amyloride, tr aldosterone antagonist diuretics such as spironolactone, epilenone et al; 2) β -adrenaline blockers such as acebutolol, atenolol, betaxolol, bevanti bopindolol, carteolol, carvedilol, celiprolol, esmolol, indenolol, metaprolol, nadolol, nebivolol, penbutolol, pindolol, probanlol, sotalol, tertatolol, tili calcium channel blockers such as amlodipine, aranidipine, azelnidipine, bamidipine, benidipine, bepridil, cinaldipine, clevidipine, diltiazem, efonid gallopamil, isradipine, lacidipine, lemidipine, lercanidipine, nicardipine, nifedipine, nilvadipine, nimodipine, nisoldipine, nitrendipine, manidipine, p et al; 4) aniotensin convertina enzyve inhibitors such as benazepril. captopril. cilazapril. delapril. enalapril. fosinopril. imidapril. rosinopril. moexi

Figure 13: Zofenopril - free text search query

4 Conclusions and Challenges

The patents case study sets up the grounds where to put together several technologies in order to come up with a useful platform for multilingual patent retrieval system. The main challenges addressed in the patent prototype are a) to translate semantically enriched patent documents, including the original markup, b) to design the mechanisms to enable the multilingual indexing and retrieval of the patents, c) to define and develop a query language and the query grammar for the system and d) to set up an on-line application for retrieval of patent document that serves as a testbed of our work.

The patents prototype combines two different approaches for machine translation, semantic annotations and retrieval techniques. The translation of the patent documents is mainly based on statistical machine translation techniques, although certain hybridization with rule-based systems (the GF) improves the quality of the translation, as discussed in WP5. One of the challenges in this task was to come up with a mechanism to translate the semantics of the source texts to the target languages. What remains as a future challenge is the use of these annotations to still increase either the accuracy of the annotations or the quality of the translations.

The GF has been proved an efficient way of generating the SPARQL queries, as if it was “Yet Another Query Language”. In other words, it allows to translate a natural language query from the user’s language to SPARQL, which makes the system accessible to a broader community rather than just skilled users. This automation facilitates also the interoperability between the query grammar and the ontologies and speeds up the development and maintenance of the querying subsystem.

Finally, the patent prototype is not comparable with the interfaces exposed by the European Patent Office, namely because they were conceived for different purposes. Nonetheless, the MOLTO patents prototype demonstrates that a patents retrieval system that addresses multilingualism by means of automatic translation techniques is commercially viable. This topic will be further discussed as part of WP10.

References

- [AFG⁺06] S. Armstrong, M. Flanagan, Y. Graham, D. Groves, B. Mellebeek, S. Morrissey, N. Stroppa, and A. Way. MaTrEx: Machine Translation Using Examples. In *TC-STAR OpenLab on Speech Translation*, Trento, Italy, 2006.
- [CCD10] Marie Candito, Benoit Crabbé, and Pascal Denis. Statistical French dependency parsing: Treebank conversion and first results. In *The seventh international conference on Language Resources and Evaluation (LREC)*, Valletta, Malta, 2010.
- [CJ05] Eugene Charniak and Mark Johnson. Coarse-to-Fine N-best Parsing and Max-Ent Discriminative Reranking. In *Proc. 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2005.
- [CNDA10] Marie Candito, Joakim Nivre, Pascal Denis, and Enrique Henestroza Anguiano. Benchmarking of Statistical Dependency Parsers for French. In *Proceedings 23rd International Conference on Computational Linguistics (COLING): Poster volume*, pages 108—116, Beijing, China, 2010.
- [GM10a] Jesús Giménez and Lluís Màrquez. Asiya: An Open Toolkit for Automatic Machine Translation (Meta-)Evaluation. *The Prague Bulletin of Mathematical Linguistics*, (94):77–86, 2010.
- [GM10b] Jesús Giménez and Lluís Màrquez. Linguistic measures for automatic machine translation evaluation. *Machine Translation*, 24(3-4):209–240, December 2010.
- [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.
- [Lin98] Dekang Lin. Dependency-based Evaluation of MINIPAR. In *Proc. Workshop on the Evaluation of Parsing Systems*, 1998.
- [MB04] Sougata Mukherjea and Bhuvan Bamba. BioPatentMiner: an Information Retrieval System for Biomedical Patents. In *Proceedings of the Thirtieth international conference on Very large data bases - Volume 30*, VLDB '04, pages 1066–1077, 2004.
- [MOL10] MOLTO. D4.1 Knowledge Representation Infrastructure, 2010.
- [MOL11] MOLTO. D5.1 Description of the final collection of corpora, September 2011.
- [MOL12a] MOLTO. D5.2 Description and Evaluation of the Combination Prototypes , March 2012.

- [MOL12b] MOLTO. D5.3 WP5 final report: statistical and robust MT, 2012.
- [MOL12c] MOLTO. D7.1 Patent MT and Retrieval Prototype Beta, January 2012.
- [MOL12d] MOLTO. D7.2 Patent MT and Retrieval Prototype, September 2012.
- [MOL13] MOLTO. D8.3 Translation and Retrieval System for Museum Object Descriptions, April 2013.
- [NHN⁺07] Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95–135, 2007.
- [PBTK06] Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. Learning accurate, compact, and interpretable tree annotation. In *21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, pages 433–440, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.
- [PK07] Slav Petrov and Dan Klein. Improved Inference for Unlexicalized Parsing. In *Proceedings Human Language Technologies (HLT)*, pages 404–411. Association for Computational Linguistics, April 2007.
- [Ran09] Aarne Ranta. The GF resource grammar library. *Linguistic Issues in Language Technology*, 2(1), 2009.
- [TWS10] John Tinsley, Andy Way, and Páraic Sheridan. PLuTO: MT for Online Patent Translation. In *Proceedings of the 9th Conferences of the Association for Machine Translation in the Americas (AMTA 2010)*, 2010.
- [VFC05] David Vallet, Miriam Fernández, and Pablo Castells. An ontology-based information retrieval model. In *In ESWC*, pages 455–470. Springer, 2005.

A Example Showing the Translation Steps of an Excerpt of Patent

The excerpt in the following example belongs to the first claim-text of the English section for claims of the patent number EP1330442B1.

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

```
The use of a compound of the formula:
<chemistry id="chem0028" num="0028"><img id="ib0030" file="imgb0030.tif" wi="55" he="37"
img-content="chem" img-format="tif"/>
</chemistry>
or isomers i.e. geometric, optical, entianomeric, diastereomeric, epimeric,
stereoisomeric, tautomeric, conformational, or anomeric forms, salts, solvates and
chemically protected forms thereof, in the preparation of a medicament for inhibiting
the activity of PARP, wherein:
<claim-text>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;
</claim-text>
<claim-text>R<sub>C</sub> is -CH<sub>2</sub>-R<sub>L</sub>, where R<sub>L</sub> is a C<sub>5-20</sub>
aryl group, 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, acyl, carboxy, ester, amido, amino,
sulfonamido, acylamido, ureido, acyloxy, thiol, thioether, sulfoxide and sulfone;
and
</claim-text>
<claim-text>R<sub>N</sub> is hydrogen.</claim-text>
```

2. Annotated text with UTF-8 encoding.

```
The use of a compound of the formula:
<chemistry id="chem0028" num="0028"><img id="ib0030" wi="55" img-format="tif" file="
imgb0030.tif" img-content="chem" he="37"/>
</chemistry>
or isomers i.e. geometric, optical, entianomeric, diastereomeric, epimeric,
stereoisomeric, tautomeric, conformational, or anomeric forms, salts, solvates and
chemically protected forms thereof, in the preparation of a medicament for inhibiting
the activity of
<AnatomicalStructure inst="umls/id/C1538577" class="semanticnetwork/id/T017">PARP
</AnatomicalStructure>
, wherein:
<claim-text>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;
</claim-text>
<claim-text>R<sub>C</sub> is -CH<sub>2</sub>-R<sub>L</sub>, where R<sub>L</sub> is a C<sub>5-20</sub> aryl group, 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, acyl, carboxy, ester, amido, amino,
sulfonamido, acylamido, ureido, acyloxy, thiol, thioether, sulfoxide and sulfone;
and
</claim-text>
<claim-text>R<sub>N</sub> is hydrogen.</claim-text>
```

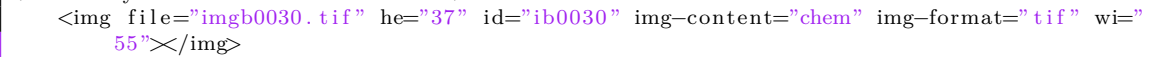
3. Marked text ready for translation.

The use of a compound of the formula : $_w$ $_z$ or isomers i.e. geometric , optical, entianomeric , diastereomeric , epimeric , stereoisomeric , tautomeric , conformational, or anomeric forms , salts, solvates and chemically protected forms thereof , in the preparation of a medicament for inhibiting the activity of $_f$ $_z$ PARP $_f$ $_z$, wherein : $_f$ $_z$ 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 ₁₋₇ alkyl, C ₃₋₂₀ heterocyclyl and C ₅₋₂₀ aryl; $_f$ $_z$ R C is $\text{--CH}_2\text{--R L}$, where R L is a C ₅₋₂₀ aryl group , optionally substituted with one or more substituent groups selected from C ₁₋₇ alkyl, C ₅₋₂₀ aryl, C ₃₋₂₀ heterocyclyl, halo, hydroxy, ether, nitro , cyano, acyl, carboxy, ester , amido , amino, sulfonamido, acylamido, ureido , acyloxy, thiol, thioether, sulfoxide and sulfone; and $_f$ $_z$ R N is hydrogen .

4. Translated patent text with the marks for reconstruction.

Verwendung einer Verbindung der Formel : $_f$ $_w$ $_z$ oder Isomere , d.h. geometrischen , optical, entianomeric , diastereomere , epimeric , stereoisomeren , tautomeren , conformational, oder anomeren bildet , salts, Solvate und chemisch geschützte Formen davon , bei der Herstellung eines Arzneimittels zur Hemmung der Aktivität von $_f$ $_z$ PARP $_f$ $_z$, wobei : $_f$ $_z$ A und B gemeinsam für einen kondensierten aromatischen Ring , gegebenenfalls substituiert mit einem oder mehreren Substituentengruppen , ausgewählt aus halo, Nitro , hydroxy, ether, thiol, thioether, amino, C ₁₋₇ alkyl, C ₃₋₂₀ Heterocyclyl und C ₅₋₂₀ aryl; $_f$ $_z$ R C $\text{--CH}_2\text{--R L}$, wobei R ein C ₅₋₂₀ Arylgruppe , gegebenenfalls substituiert mit einem oder mehreren Substituentengruppen , ausgewählt aus C ₁₋₇ alkyl, C ₅₋₂₀ aryl, C ₃₋₂₀ heterocyclyl, halo, hydroxy, ether, Nitro , cyano, acyl, carboxy, , Amido amino, sulfonamido, acylamido, Ureido , acyloxy, thiol, thioether, Sulfoxid und sulfone; und $_f$ $_z$ N R Wasserstoff ist .

5. Translated patent text reconstructed using the original document EP1330442B1.

Verwendung einer Verbindung der Formel :
<chemistry id="chem0028" num="0028">
 </chemistry>
oder Isomere , d.h. geometrischen , optical, entianomeric , diastereomere , epimeric , stereoisomeren , tautomeren , conformational, oder anomeren bildet , salts, Solvate und chemisch geschützte Formen davon , bei der Herstellung eines Arzneimittels zur Hemmung der Aktivität von
<AnatomicalStructure class="semanticnetwork/id/T017" inst="umls/id/C1538577"> PARP
</AnatomicalStructure>
, wobei :
<claim-text> A und B gemeinsam für einen kondensierten aromatischen Ring ,
gegebenenfalls substituiert mit einem oder mehreren Substituentengruppen , ausgewählt
aus halo, Nitro , hydroxy, ether, thiol, thioether, amino, C ₁₋₇ alkyl, C ₃₋₂₀
Heterocyclyl und C ₅₋₂₀ aryl;
</claim-text>
<claim-text> R C $\text{--CH}_2\text{--R L}$, wobei R ein C ₅₋₂₀ Arylgruppe , gegebenenfalls substituiert
mit einem oder mehreren Substituentengruppen , ausgewählt aus C ₁₋₇ alkyl, C ₅₋₂₀ aryl
, C ₃₋₂₀ heterocyclyl, halo, hydroxy, ether, Nitro , cyano, acyl, carboxy, , Amido
amino, sulfonamido, acylamido, Ureido , acyloxy, thiol, thioether, Sulfoxid und
sulfone; und
</claim-text>
<claim-text> N R Wasserstoff ist. </claim-text>

B Patent Translator API - Specification

This appendix contains the instructions to run the patent translation systems from any another application. The patent translator API is a set of scripts that can translate a (set of) full XML patent(s). It returns back the same patent(s) having the requested translations included into the XML schema. In addition, it is also possible to translate *raw* files, which enables the chance to translate *raw* text from an input box in the uer interface, such as for instance the GF cloud.

All the libraries and scripts are at available in the svn and they have been installed in the UGOT server, where all the dependencies with additional software (moses, tokenizers, filtering, etc.) have been already configured. The output of the scripts is given in the following folder: `/tmp/hybriddemo/<random>` , in order to be able to process multiple translation request and to avoid problems due remote user's writing permissions.

B.1 Translation of *raw* text

This script translates the a raw input file in the source language and translates it to the target language.

The translated text is saved as `<tmp_folder>/output/<input_file>.tmp.trad`, where the `tmp_folder` is given in the `STDOUT` and the `<input_file>` is the name of the file provided as input parameter.

Usage: `./translateRAW.sh source_language target_language system input_file`
where,

- `source_language` is one of "en","de","fr". Lowercase is required.
- `target_language` is one of "en","de","fr", and different from `source_language`. Lowercase is required.
- `system` is one of "smt","hybrid". Lowercase is required.
- `input_file` is a single raw file, one sentence per line. Full path to the file must be provided. It will be tokenized and translated in a line by line basis.

Output: It returns in the `STDOUT` the folder where the translation is located.

Example: `./translateRAW.sh en de smt /soft/patdoctranslator/example/patsA61P.raw`
`2> /dev/null`

Example output: `/tmp/hybriddemo/577`

Example translated file: `/tmp/hybriddemo/577/output/patsA61P.raw.tmp.trad`

B.2 Translation of *XML* patent documents

This script can translate a set of any number of XML patent documents (they must follow the EPO XML specification for patents).

Each translated file is saved as `<tmp_folder>/output/combined.<input_file_name>`, where the `tmp_folder` is given in the `STDOUT` and the `<input_file>` is the name of the XML file provided as input parameter.

Usage: `./translateXML.sh source_language target_language system <xml_files>`
where,

- `source_language` is one of “en”, “de”, “fr”. Lowercase is required.
- `target_language` is one of “en”, “de”, “fr”, and different from `source_language`. Lowercase is required.
- `system` is one of “smt”, “hybrid”. Lowercase is required.
- `<xml_files>` is any number greater than zero of additional arguments of XML files (full path must be provided).

Output: It returns in the `STDOUT` the folder where the translations are located.

Example: `./translateXML.sh en de smt /soft/patdoctranslator/example/EP2033635B1.xml`
`2> /dev/null`

Example output: `/tmp/hybriddemo/23642`

Example translated file: `/tmp/hybriddemo/23642/output/combined.EP2033635B1.xml`

C Ontologies in the Biomedical Domain

Several ontologies from the biomedical domain that are aligned for the patent retrieval prototype. This is the complete list of all datasets that are loaded in the semantic repository, 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.owl and proton classes.

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

measure-unit-classes.owl: measurement units ontology

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 G-proteins coupled receptors instances.

pharma-params-instances.nt: FDA pharmaceutical parameters.

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_tripplles.n3: triples from the new annotations in the patent documents.

D Semantic Annotations - Final Types

Hereby we provide the list of semantic annotations that were kept for the final version of the prototype.

Annotation	Semantics
ActiveIngredient	The active ingredients recognized by the FDA. Patent and generic FDA-approved drugs may have the same active ingredients. http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#ActiveIngredient
AnatomicalStructure	A normal or pathological part of the anatomy or structural organization of an organism. http://linkedlifedata.com/resource/semanticnetwork/id/T017
Applicant	This is the company who has applied to the FDA for approval. Usually synonymous with "manufacturer." http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#Applicant
ApplicationDate	EPO patent application date(extracted from the patent's metadata). http://proton.semanticweb.org/2006/05/patents#ApplicationDate
ApplicationNumber	EPO patent application number(extracted from the patent's metadata). http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#ApplicationNumber
DiseaseOrDysfunction	A disordered process, activity, or state of the organism as a whole, of a body system or systems, or of multiple organs or tissues. Included here are normal responses to a negative stimulus as well as pathologic conditions or states that are less specific than a disease. Pathologic functions frequently have systemic effects. http://linkedlifedata.com/resource/semanticnetwork/id/T046
DosageForm	This is form in which the approved strength can be administered. Approved drugs will be specific to a strength and dosage form, although one approval may include several dosage/strength/route of administration combinations. http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#DosageForm
Drug	Usually the commercial name in the case of a trade drug. If generic, may be only the active ingredient. http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#FDA_DrugName
Measurement	General type for variety of measurements. http://proton.semanticweb.org/2006/05/measure
PatentNumber	EPO patent number(extracted from the patent's metadata). http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#PatentNumber
PharmaParam	A pharmacological parameter indicating a level or boundary. http://proton.semanticweb.org/2006/05/measure#PharmaParam
PublicationDate	EPO patent publication date(extracted from the patent's metadata). http://proton.semanticweb.org/2006/05/patents#PublicationDate
Receptor	A specific structure or site on the cell surface or within its cytoplasm that recognizes and binds with other specific molecules. These include the proteins on the surface of an immunocompetent cell that binds with antigens, or proteins found on the surface molecules that bind with hormones or neurotransmitters and react with other molecules that respond in a specific way. http://linkedlifedata.com/resource/semanticnetwork/id/T192
RouteOf Administration	The method by which a drug is taken, i.e. oral vs. injection. http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#RouteOfAdministration

Table 10: Semantic annotation types for the final prototype

E Biomedical Patents - Query Patterns

The patents query language covers the set of query topics shown in Table 11. The grammar was augmented in order to cover better the relations in the ontologies. Details about it are given in Section 2.3. Below is the list of all generic queries that can be asked to the system.

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 11: The biomedical patent query topics

```

what is the information about ACTIVE_INGREDIENT
what are the patent numbers
what is the approval date of the patent for DRUG
what is the patent with the application number APP_NUMBER
what is the information about APPLICANT
what is the application date of PATENT
what is the application number for PATENT
what is the approval date of the applicant APPLICANT
what is the strength of drugs that contain CHEMICAL_SUBSTANCE
what are the drugs with the active ingredient ACTIVE_INGREDIENT
what are the active ingredients of DRUG
what are the drugs that mention APPLICANT
what are the dosage forms of DRUG
what is the information about DRUG
what are the drugs with the market MARKET
what are the names of DRUG
what are the drugs with the route of administration ROUTE_OF_ADMINISTRATION
what is the strength of DRUG
what are the drugs with the therapeutic equivalence code TE_CODE
what is the therapeutic equivalence code of DRUG
what is the expiration date of the patent for DRUG
what are the markets of DRUG
PATENT

```

what is the expiration date of the application number APP_NUMBER
 what is the expiration date of PATENT
 what is the information about ROUTE_OF_ADMINISTRATION
 what are the patents that mention APPLICANT
 what are the patents that mention DRUG
 what are the patents that mention ACTIVE_INGREDIENT
 what are the patents that mention ROUTE_OF_ADMINISTRATION
 what are the patents that mention APPLICANT and APPLICANT
 what are the patents that mention APPLICANT and DRUG
 what are the patents that mention APPLICANT and ACTIVE_INGREDIENT
 what are the patents that mention APPLICANT and ROUTE_OF_ADMINISTRATION
 what are the patents that mention DRUG and APPLICANT
 what are the patents that mention DRUG and DRUG
 what are the patents that mention DRUG and ACTIVE_INGREDIENT
 what are the patents that mention DRUG and ROUTE_OF_ADMINISTRATION
 what are the patents that mention ACTIVE_INGREDIENT and APPLICANT
 what are the patents that mention ACTIVE_INGREDIENT and DRUG
 what are the patents that mention ACTIVE_INGREDIENT and ACTIVE_INGREDIENT
 what are the patents that mention ACTIVE_INGREDIENT and ROUTE_OF_ADMINISTRATION
 what are the patents that mention ROUTE_OF_ADMINISTRATION and APPLICANT
 what are the patents that mention ROUTE_OF_ADMINISTRATION and DRUG
 what are the patents that mention ROUTE_OF_ADMINISTRATION and ACTIVE_INGREDIENT
 what are the patents that mention ROUTE_OF_ADMINISTRATION and ROUTE_OF_ADMINISTRATION

Each of them has several possible verbalizations, for example for show me the information about AMPICILLIN these are:

what is the information about AMPICILLIN
 information about AMPICILLIN
 the information about AMPICILLIN
 give me the information about AMPICILLIN
 show me the information about AMPICILLIN
 what information do you have about AMPICILLIN
 what information can I get about AMPICILLIN
 all information about AMPICILLIN
 all about AMPICILLIN
 AMPICILLIN

The French versions of the above queries are:

quelle est l' information à propos de AMPICILLIN
 information à propos de AMPICILLIN

l' information à propos de AMPICILLIN
montre l' information à propos de AMPICILLIN
quelle information avez à propos de AMPICILLIN vous
quelle information peux obtenir à propos de AMPICILLIN je
toute de l' information à propos de AMPICILLIN
tout à propos de AMPICILLIN
AMPICILLIN

And the German verbalizations of the above query are:

was ist die Information über AMPICILLIN
Information über AMPICILLIN
die Information über AMPICILLIN
zeige mir die Information über AMPICILLIN
welche Information habt ihr über AMPICILLIN
welche Information kann ich über AMPICILLIN bekommen
alle Information über AMPICILLIN
alles über AMPICILLIN
AMPICILLIN

F Patent Retrieval Databases Roadmap

A complete list of the patent retrieval databases was given in the previous Deliverable D7.2. For the readability of this document, we have included this appendix with a shorter selection of them. They constitute the concepts contained in the ontologies that can be also found in the patent documents indexed in the system.

Drug	#docs	Drug	#docs
ACETIC ACID	1829	PREDNISONE	318
SODIUM CHLORIDE	1434	IBUPROFEN	317
TALC	1366	AMPICILLIN	295
INSULIN	1042	INDOMETHACIN	279
SODIUM BICARBONATE	819	MITOXANTRONE	272
PENICILLIN	765	MERCAPTOPYRINE	257
MAGNESIUM SULFATE	619	CYTARABINE	243
AMMONIUM CHLORIDE	597	PREDNISOLONE	234
DIMETHYL SULFOXIDE	525	NAPROXEN	231
STERILE WATER	509	ESTRADIOL	229
ADENOSINE	499	LOVASTATIN	227
CYCLOPHOSPHAMIDE	490	SIMVASTATIN	226
CISPLATIN	449	IFOSFAMIDE	225
DEXAMETHASONE	442	HYDROXYUREA	222
FLUOROURACIL	439	POTASSIUM CHLORIDE	222
ETOPOSIDE	404	FOLIC ACID	219
PACLITAXEL	379	VITAMIN D	207
TAXOL	360	TAXOTERE	205
MITOMYCIN	355	TESTOSTERONE	203
CARBOPLATIN	354	FLUTAMIDE	201

Table 12: Drugs most mentioned in patents.

Active Ingredient	#docs	Active Ingredient	#docs
CALCIUM	2573	COPPER	722
ALCOHOL	2391	LACTIC ACID	698
AMINO ACIDS	1777	SODIUM SULFATE	679
SODIUM CHLORIDE	1434	GLYCERIN	662
TALC	1366	CALCIUM CARBONATE	660
MANNITOL	1362	BIOTIN	621
CITRIC ACID	1208	MAGNESIUM SULFATE	619
SORBITOL	1206	AMMONIUM CHLORIDE	597
GLYCINE	1160	SODIUM ACETATE	586
SULFUR	1033	DIMETHYL SULFOXIDE	525
PHOSPHORIC ACID	1018	SODIUM CITRATE	508
TYROSINE	979	ADENOSINE	499
GLUTAMINE	955	CYCLOPHOSPHAMIDE	490
PROTEASE	903	SODIUM PHOSPHATE	480
SODIUM CARBONATE	878	CISPLATIN	449
TARTARIC ACID	821	DEXAMETHASONE	442
SODIUM BICARBONATE	819	FLUOROURACIL	439
DEXTROSE	804	ETOPOSIDE	404
ASCORBIC ACID	802	ASPIRIN	384
UREA	774	PACLITAXEL	379

Table 13: Active ingredients most mentioned in patents.

Drug	Administration	#docs	Drug	Administration	#docs
SODIUM CHLORIDE	INJECTION	1220	PACLITAXEL	INJECTION	312
INSULIN	INJECTION	802	MITOMYCIN	INJECTION	310
SODIUM BICARBONATE	INJECTION	630	CARBOPLATIN	INJECTION	299
PENICILLIN	ORAL	594	TAXOL	INJECTION	295
MAGNESIUM SULFATE	INJECTION	500	IBUPROFEN	ORAL	291
AMMONIUM CHLORIDE	INJECTION	470	PREDNISONE	ORAL	282
CYCLOPHOSPHAMIDE	ORAL	439	FLUOROURACIL	TOPICAL	274
CYCLOPHOSPHAMIDE	INJECTION	420	INDOMETHACIN	ORAL	257
ADENOSINE	INJECTION	399	MITOXANTRONE	INJECTION	249
DEXAMETHASONE	ORAL	388	MERCAPTOPYRINE	ORAL	237
CISPLATIN	INJECTION	385	CYTARABINE	INJECTION	220
FLUOROURACIL	INJECTION	381	NAPROXEN	ORAL	219
DEXAMETHASONE	INJECTION	380	AMPICILLIN	ORAL	217
ETOPOSIDE	ORAL	357	PREDNISOLONE	ORAL	211
ETOPOSIDE	INJECTION	353	LOVASTATIN	ORAL	210

Table 14: Drug names and routes of administration that are most mentioned in documents.

Drug	Dosage Form	#docs	Drug	Dosage Form	#docs
TALC	POWDER	1052	CYCLOPHOSPHAMIDE	TABLET	222
SODIUM CHLORIDE	INJECTABLE	722	CISPLATIN	INJECTABLE	216
DIMETHYL SULFOXIDE	SOLUTION	521	ADENOSINE	INJECTABLE	215
STERILE WATER	LIQUID	465	INDOMETHACIN	SUSPENSION	212
DEXAMETHASONE	SOLUTION	432	ETOPOSIDE	INJECTABLE	206
INSULIN	INJECTABLE	423	TAXOL	INJECTABLE	192
FLUOROURACIL	SOLUTION	418	CYCLOSPORINE	SOLUTION	190
SODIUM BICARBONATE	INJECTABLE	370	PACLITAXEL	INJECTABLE	182
PREDNISONE	SOLUTION	309	NAPROXEN	SUSPENSION	178
MAGNESIUM SULFATE	INJECTABLE	260	CARBOPLATIN	INJECTABLE	172
AMMONIUM CHLORIDE	INJECTABLE	253	DEXAMETHASONE	TABLET	168
IBUPROFEN	SUSPENSION	248	MITOMYCIN	INJECTABLE	166
CYCLOPHOSPHAMIDE	INJECTABLE	243	ETOPOSIDE	CAPSULE	164
FLUOROURACIL	INJECTABLE	232	MITOXANTRONE	INJECTABLE	159
DEXAMETHASONE	INJECTABLE	230	BALANCED SALT	SOLUTION	145

Table 15: Drug names and dosage forms that are most mentioned in documents.

Drug	Active Ingredient	#docs
ABELCET	AMPHOTERICIN B	2
ABILIFY	ARIPIPRAZOLE	8
ABRAXANE	PACLITAXEL	28
ACARBOSE	ACARBOSE	152
ACCOLATE	ZAFIRLUKAST	7
ACCUPRIL	QUINAPRIL HYDROCHLORIDE	3
ACCUTANE	ISOTRETINOIN	5
ACEBUTOLOL HYDROCHLORIDE	ACEBUTOLOL HYDROCHLORIDE	7
ACEON	PERINDOPRIL ERBUMINE	5
ACETAMINOPHEN	ACETAMINOPHEN	161
ACETAZOLAMIDE	ACETAZOLAMIDE	50
ACETOHEXAMIDE	ACETOHEXAMIDE	53
ACETYLCYSTEINE	ACETYLCYSTEINE	61
ACTH	CORTICOTROPIN	11
ACTOS	PIOGLITAZONE HYDROCHLORIDE	4
ACYCLOVIR	ACYCLOVIR	109
ACYCLOVIR	ACYCLOVIR SODIUM	2
ACYCLOVIR SODIUM	ACYCLOVIR SODIUM	2
ADAGEN	PEGADEMASE BOVINE	3
ADALAT	NIFEDIPINE	4
ADENOSINE	ADENOSINE	499
ADRUCIL	FLUOROURACIL	9
ADVICOR	NIACIN	8
ADVICOR	LOVASTATIN	10
ADVIL	IBUPROFEN	3
AEROBID	FLUNISOLIDE	2
AEROLATE SR	THEOPHYLLINE	1
AGENERASE	AMPRENAVIR	15

Table 16: Drug names, the active ingredients of which are mentioned in documents.