MOLTO

Multilingual Online Translation

Non multa, sed multum

Contract No.: FP7-ICT-247914
Project full title: MOLTO - Multilingual Online Translation
Deliverable: D7.1 Patent MT and Retrieval Prototype Beta
Security (distribution level): Public, regular publication
Contractual date of delivery: M21
Actual date of delivery: January, 2012
Type: Prototype
Status & version: final, v1.5

Author(s): Milen Chechev[2], Ramona Enache[3], Cristina España-Bonet[1], Meritxell Gonzàlez[1], Lluiś Màrquez[1], Borislav Popov[2], Aarne Ranta[3]
Task responsible: UPC[1]
Other contributors: Ontotext[2], UGOT[3]

ABSTRACT

This document is the written report of the first deliverable corresponding to WP7, *Case Study: Patents*. It describes the preliminary prototype for patent translation and retrieval.

First, there is a general overview of the workpackage and we briefly summarise the scenarios considered within the prototype. Then, we give the general layout of the prototype architecture, the demonstrator interface and the technologies integrated in the prototype. Finally, we summarise the current status of the workpackage and the future directions for the final prototype.

Contents

1 Introduction

This document is the first deliverable corresponding to WP7, *Case Study: Patents*. It describes the preliminary prototype for Patent Translation and Retrieval.

Nowadays, there are five main patent offices around the world: the United States, Europe, China, Korea and Japan. These offices manage a huge amount of documents describing the patented inventions. There is a clear need to exchange the information related to such inventions, either for carrying out the legal tasks characteristic of the patent offices, or for building systems able to access, search for and translate the patent data and make them available to the international community. However, these offices use different languages for their written documents and it is not possible to undertake the task of translating all the document using human resources (e.g., due the outsize of the databases or the update frequency of the patent documents). Therefore, we consider that this is an interesting case study in which to apply the technologies developed within the MOLTO project.

The mission of the MOLTO project is to enable multilingual translation with high quality (grammatically and stylistically) and sufficient level of speed and automation for real-time translation tasks. In the case study of this workpackage, we aim to create a prototype for MT and retrieval of patents in the biomedical and pharmaceutical domains, allowing translation of patent abstracts and claims and exposing several cross-language retrieval paradigms on top of them.

Our translation target languages are English, French and German, since these are the official languages of the European Patent Office (EPO). According to the European Patent Conventions, every patent application shall be filled, at least, in one of the official languages. Moreover, the specifications of the European patents shall be published in the language of the proceedings and shall include a translation of the claims to the other two official languages.

The retrieval system in MOLTO uses a semantic infrastructure acting as a central multi-paradigm index for upper-level conceptual models and domain ontologies, knowledge bases, patents content and metadata; and providing NL-based retrieval. There is a database of legacy documents ready to use, but no ready-made ontology is available with sufficient coverage of the domain.

Broadly, the prototype will include both technologies mentioned above, and will be tested and evaluated according to general criteria in terms of usability and translation quality. Translation quality can be assessed on the grounds of human evaluation and a combination of automatic metrics (i.e. BLEU [PRWZ02] or Asiya [GM10]). In order to assess usability, we will examine the feasibility of the prototype as part of a commercial patent retrieval system.

The workpackage is tightly related to WP5, *Statistical and Robust Translation*, and WP4, *Knowledge Engineering*. The engines for the translation systems are built within WP5, whereas the ones for the retrieval systems and its interoperability with the Grammatical Framework (GF[1]) are built in WP4. Both technologies are integrated into the patents case study prototype described herein.

This document is organised as follows. In the following, Section 2 details the scenarios for the patents case study. Section 3 gives a general overview of the patent prototype and the details for the current preliminary systems. Finally, Section 4 summarises the current status of the workpackage and the future directions for the final prototype.

2 General Use Cases for the Patents Case Study

The patents case study comprises two basic scenarios: the online patent retrieval and the patent translation. In this prototype we tackle these two scenarios separately, as shown in Figure 1, even though they can be viewed as a join multilingual patent retrieval paradigm. In the future, MOLTO will study how to automate the reciprocal inputs between the two processes, i.e., the annotation of translations and the translation of semantically annotated documents.

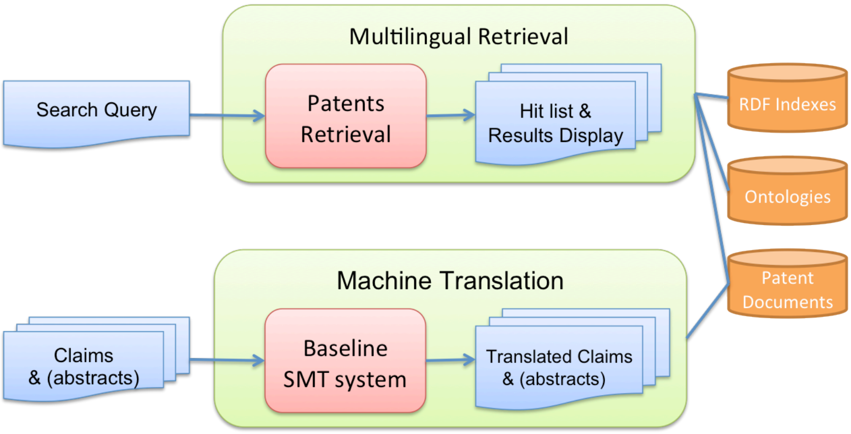


Figure 1: Scenarios at the patent case study

From a general perspective, two user roles may be defined in this case study: end-users looking for information related to the patents and editors adding new patent documents to a hypothetical repository. Figure 2 shows the general workflow for both user roles. The green path corresponds to the end-user, whereas the blue one corresponds to a patent editor/translator workflow. The red and orange boxes in this Figure are, respectively, tasks not implemented or partially implemented in the current prototype. The dashed arrows show the path not planned to be implemented in the prototype, but that we have considered in terms of usability and feasibility of the system for future purposes. Namely, this path corresponds to the interaction between a patent editor and the system. The prototype does not include the development of such a user interface, which may depend on the particularities of a hypothetical information system. Nevertheless, in MOLTO we will

[1] http://www.grammaticalframework.org/

design a pipeline to favour and facilitate any future development in this line. Finally, the boxes in the centre of the Figure correspond to the specific tasks for the patent retrieval and translation systems, i.e. classification of documents and queries, storage, search, retrieval and translation.

Patents Structure. The files associated to every patent, normalised to an XML format, contain the terms of the patent and the bibliographic data. The standardised fields include dates, countries, languages, references, author names and companies as well as rich subject classifications. Moreover, every patent has a title, a description, an abstract with a short and general summary and a series of claims. The Deliverable 5.1 [MOL11] gives a detailed description of the patent documents.

Patents Retrieval. In the patent retrieval scenario, end-users have access, in their language, to some information that may be originally produced in other languages. On the one hand, the patent documents are classified in multiple indexes according the to information about the patent (e.g. the bibliographic data of the patent or the content of the claims), and the language they are written. On the other hand, the end-user searches for patents matching some criteria. In MOLTO, such criteria are written in the user's own language, independently of the language of the source documents, and it is translated into a relational representation between the terms of the query and the content of the patent indexes. The online translation of NL queries is grounded on the abstract syntax representation produced by the GF [Ran11]. The current interface, described in Section 3.1, allows to query the system in English and French under a controlled language designed for the patents domain. A hit list of patents is shown to the user along with a brief answer produced in natural language (NL) in the same language of the user's query, and the set ontological concepts that matched the query. This way, the user could have an idea of the content, select any of the documents to see the whole content or the semantic annotations, or update the query to obtain further results.

Patents Translation. The patent translation scenario refers to the off-line translation of the patent's claims and abstracts written in different languages. Patents can be translated when they are added to the repository (e.g. by the editor) or when they are retrieved from the repository (e.g. by the end-user). In MOLTO we focus on the translation of patents when they are added to the repository. The SMT system, detailed in Section 3.3, has been trained over a set of provisional corpora which only provides the claims with aligned multilingual texts (i.e. the same text written in at least two of the target languages). It allows to translate a bunch of patent claims and abstract text in raw format from/to English, French and German. Given a set of original patent documents, the text content is extracted and translated to the missing languages. Then, the text is set into the documents following the same patent format.

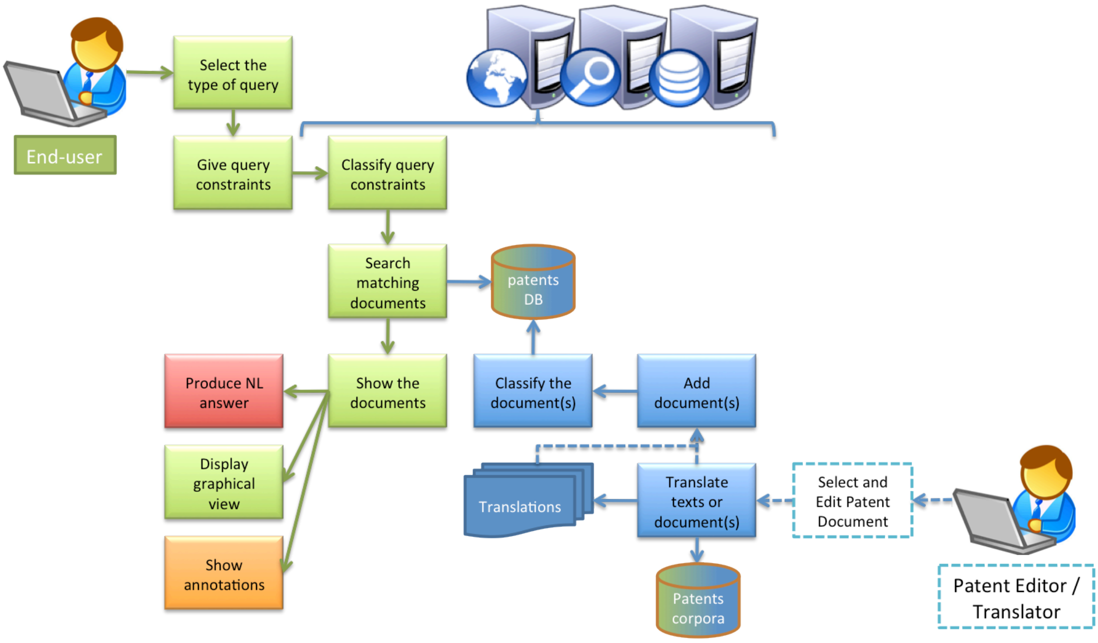


Figure 2: General workflow for user roles and scenarios

3 The Patents Beta Prototype

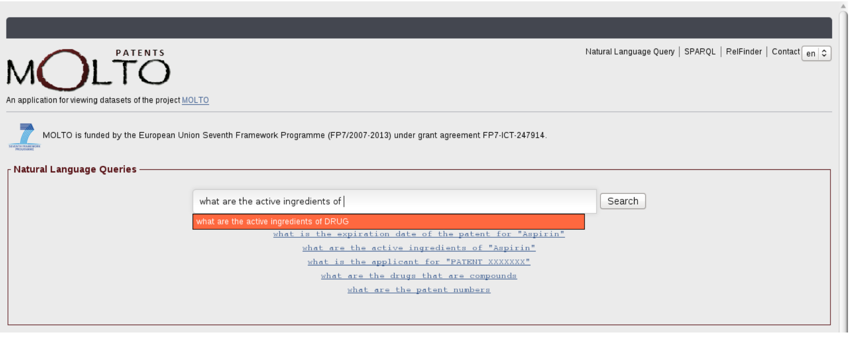
The Patents Case Study focuses on patents belonging to the biomedical and pharmaceutical domain in order to limit the scope of the problem and cope with the technological requirements. The prototype described in this document, which can be accessed at `http://molto-patents.ontotext.com/`, is a fully functional beta version in which we have set the grounds for the final prototype. In the following sections we describe the components of the online demo: the web interface, the retrieval system and the SMT baseline.

3.1 The Demo Web Interface

The demo deals with patents translated to English and French, and so are the languages of the interface, although in the final version we will include German. The patent retrieval interface exposes several query types, including the natural language (NL) and the RelFinder (RF) tool. The NL interface interprets queries written in English and French and shows up the hit list of documents, highlighting the annotations on them. The RF tool displays the relations given by the semantic resources among instances or concepts included on them.

3.1.1 The Natural Language Based Interface

The natural language interface allows the user to give the search criteria using a controlled language. The specific query grammar processes the user input. Every possible input described by the controlled language has a correspondence with a grammar rule in the GF and generates an abstract syntax tree that is translated into SPARQL[2]. Moreover, the grammar has been integrated in the interface so that it enables an autocomplete function to help the user writing queries under the controlled language supported by the grammar. The autocomplete functionality is shown in Figure 3.



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.

what are the active ingredients of

what are the active ingredients of DRUG

what is the expiration date of the patent for "Aspirin"

what are the active ingredients of "Aspirin"

what is the applicant for "PATENT XXXXXXX"

what are the drugs that are compounds

what are the patent numbers

Figure 3: The natural language query interface

The query grammar covers a set of query topics, shown in Table 1, for which we wrote a number of query examples. The initial set of query examples consisted of 131 sentences in English. Nonetheless, the current version of the grammar generates (and therefore can process) a wider spectrum of sentences. In particular it generates 591 sentences in English and 504 sentences in French. The difference between both languages is due to the specific characteristics of each language. Table 2 gives some examples of the patent queries in English and French.

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 1: The patent query topics

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

English	French
what information can I get about DRUG	*quelle information puis-je obtenir à propos de DRUG*
what are the chemical substances of DRUG	*la substance chimique de DRUG*
what are the active ingredients of DRUG	*Quels sont les ingrédients actifs de DRUG*
give me the drugs that are compounds	*montre les médicaments qui sont des compos*
what are the dosage forms of DRUG	*Quelles sont les formes posologiques de DRUG*
the drug preparations for DRUG	*quelles préparations y a-t-il*
what is the route of administration of DRUG	*quelles sont les voies d'administration de DRUG*
I want the name of a DRUG	*je veux le nom de DRUG*
what are the methods being used in PATENT	*quelles sont les méthodes de PATENT*
what are the methods of PATENT	
what is the patent number for DRUG	*Quel est le numéro de brevet pour DRUG*
when does PATENT expire?	*quand expire PATENT expire-t-il*
give me the use codes of PATENT	*montre les codes d'utilisation de PATENT*

Table 2: Patent query examples

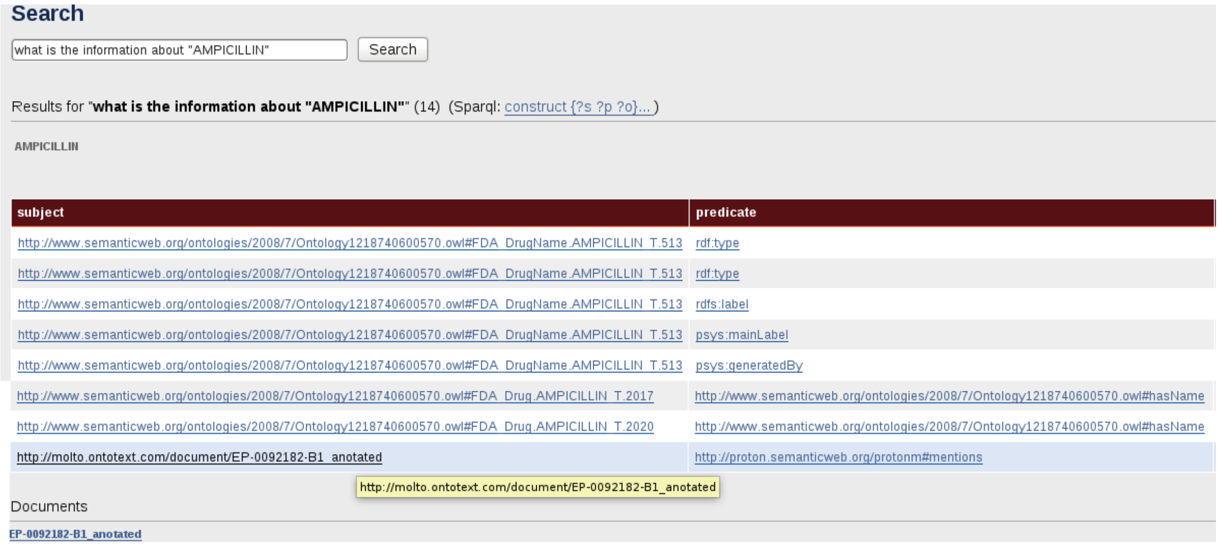
The results interface shows several data related to the user's query and its interpretation. First, the interface displays the query's translation into SPARQL language; then, the set of classes from the ontologies that match the query; and finally, the annotated documents where the data was found. For example, given the user input *"what is the information about ``AMPICILLIN''*, Figure 4 shows the results obtained for it. The translation into SPARQL is as follows:

```
construct {?s ?p ?o}
WHERE { {?s <http://www.w3.org/2000/01/rdf-schema#label> "AMPICILLIN" . ?s ?p ?o }
UNION {?o <http://www.w3.org/2000/01/rdf-schema#label> "AMPICILLIN" . ?s ?p ?o }}
```

Below the list of ontologies' items, the interface shows also a link to the semantically annotated document EP-0092182-B1, and at the bottom of the page we can also see a link to the original patent document. If we follow the former link, the interface displays the text of the patent document. The right side of the page shows the list of the semantic annotations used within the text, each one holding a colour. The interface highlights in the text, according to the colour given, those words that are related to any of the semantic items. Figure 5 shows the highlighted text for the document retrieved in the query from the example. In this example we can observe the context in which the following words are mentioned: ``Urokinase'' is an active ingredient, ``Plasminogen'' is an anatomical structure, ``lysis'' is a disease or a disfunction, ``solution'' is a dosage form, ``Acetic acid'' is a drug and ``pH'' is a measurement.

3.1.2 The RelFinder Tool

In addition to the natural language based interface, we have also included the RelFinder tool, a more graphical interface that shows up the relations among instances of the ontologies. Figure 6 shows a use example of the RelFinder tool in which the search includes the terms: ``Ampicillin'', ``Tetracycline'' and ``Acetic Acid''. The right side displays a



Search

what is the information about "AMPICILLIN" | Search

Results for "**what is the information about "AMPICILLIN"**" (14) (Sparql: construct {?s ?p ?o}...)

AMPICILLIN

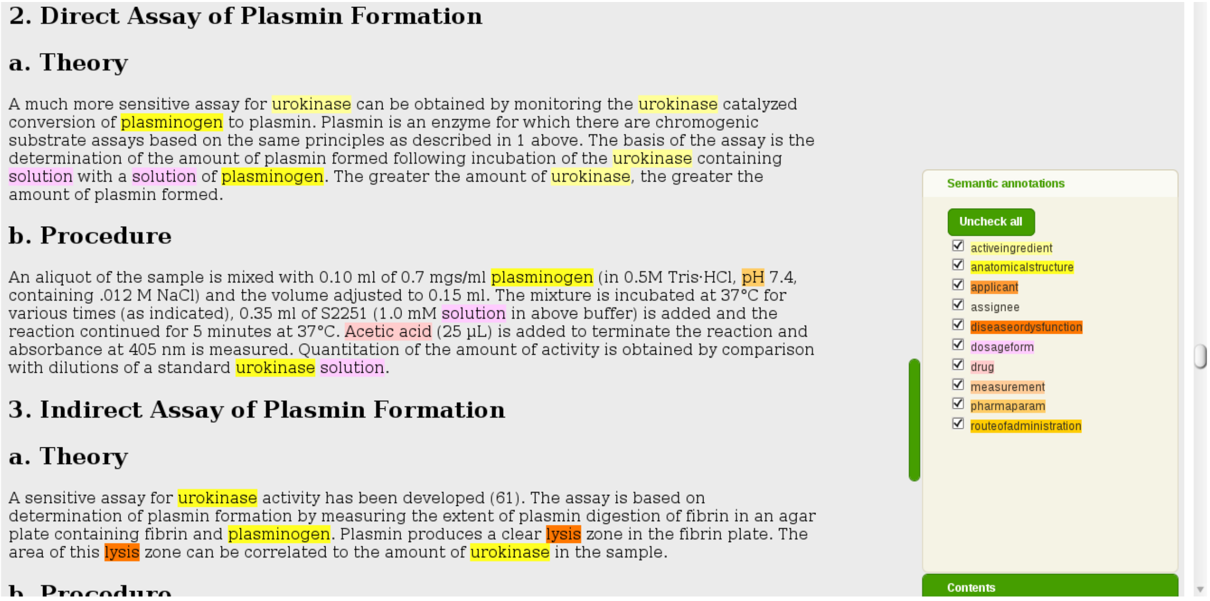
subject	predicate
http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#FDA_DrugName.AMPICILLIN_T.513	rdf:type
http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#FDA_DrugName.AMPICILLIN_T.513	rdf:type
http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#FDA_DrugName.AMPICILLIN_T.513	rdfs:label
http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#FDA_DrugName.AMPICILLIN_T.513	psys:mainLabel
http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#FDA_DrugName.AMPICILLIN_T.513	psys:generatedBy
http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#FDA_Drug.AMPICILLIN_T.2017	http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#hasName
http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#FDA_Drug.AMPICILLIN_T.2020	http://www.semanticweb.org/ontologies/2008/7/Ontology1218740600570.owl#hasName
http://molto.ontotext.com/document/EP-0092182-B1_anotated	http://proton.semanticweb.org/protonm#mentions

http://molto.ontotext.com/document/EP-0092182-B1_anotated

Documents

EP-0092182-B1_anotated

Figure 4: The interface showing the patent retrieval results



2. Direct Assay of Plasmin Formation

a. Theory

A much more sensitive assay for urokinase can be obtained by monitoring the urokinase catalyzed conversion of plasminogen to plasmin. Plasmin is an enzyme for which there are chromogenic substrate assays based on the same principles as described in 1 above. The basis of the assay is the determination of the amount of plasmin formed following incubation of the urokinase containing solution with a solution of plasminogen. The greater the amount of urokinase, the greater the amount of plasmin formed.

b. Procedure

An aliquot of the sample is mixed with 0.10 ml of 0.7 mgs/ml plasminogen (in 0.5M Tris·HCl, pH 7.4, containing .012 M NaCl) and the volume adjusted to 0.15 ml. The mixture is incubated at 37°C for various times (as indicated), 0.35 ml of S2251 (1.0 mM solution in above buffer) is added and the reaction continued for 5 minutes at 37°C. Acetic acid (25 µL) is added to terminate the reaction and absorbance at 405 nm is measured. Quantitation of the amount of activity is obtained by comparison with dilutions of a standard urokinase solution.

3. Indirect Assay of Plasmin Formation

a. Theory

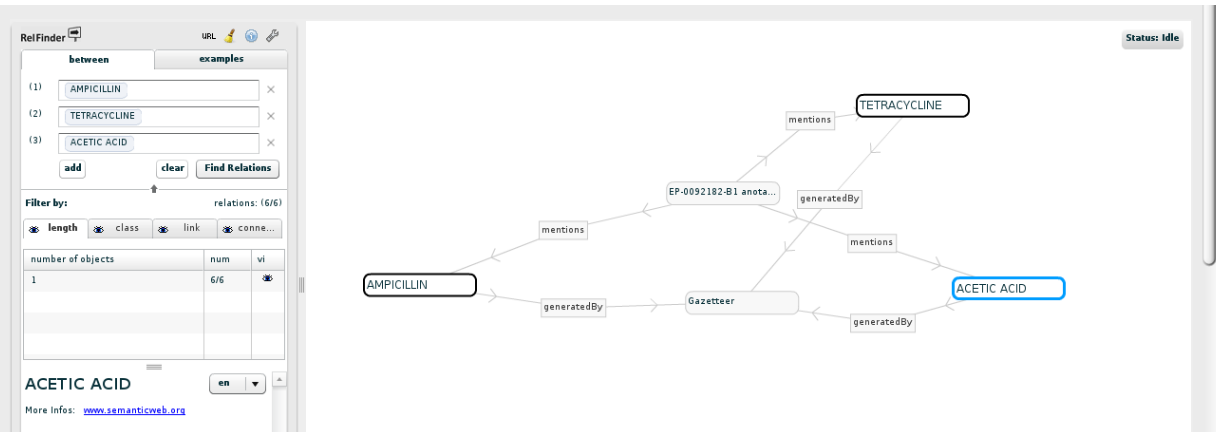
A sensitive assay for urokinase activity has been developed (61). The assay is based on determination of plasmin formation by measuring the extent of plasmin digestion of fibrin in an agar plate containing fibrin and plasminogen. Plasmin produces a clear lysis zone in the fibrin plate. The area of this lysis zone can be correlated to the amount of urokinase in the sample.

b. Procedure

Figure 5: The interface highlighting the annotations of the patent document

graph showing the connections among the terms where we can see the annotated document EP-0092182-B1 found in the natural language query example in the previous section.

The interface interaction approach is different comparing to the natural language one. The main characteristic is that the interface enables the user to write the objects that he wants to explore. The autocomplete functionality integrated in the RelFinder interface is based on the labels of all the instances in the semantic repository. Furthermore, the results are displayed as a graph view which helps in data understanding. The graph is interactive, so that the user can select the items from the results to see more information about them and even to explore the results in a similar way as in the SPARQL-based interface.



(1) AMPICILLIN
(2) TETRACYCLINE
(3) ACETIC ACID

Filter by: relations: (6/6)

number of objects	num	vi
1	6/6	👁

ACETIC ACID en

More Infos: www.semanticweb.org

Figure 6: The patent RelFinder tool interface

3.2 The Patent Retrieval System

The details of the retrieval system are given in the Deliverables 4.1, 4.2 and 4.3 provided by WP4. In gross, the engines for the patent case study are based on Exopatent[3] and KRI[4]. The KRI includes the RDFDB, the PROTON Ontology and the KRI Web UI. The former is an API that provides a remote access to the stored and structured data via JMS. The PROTON Ontology is a lightweight upper-level ontology which defines about 300 classes and 100 properties, covering most of the upper-level concepts, necessary for semantic annotation, indexing and retrieval. The KRI Web UI is a user interface that accesses OWLIM through the RDFDB layer and gives the user the possibility to browse the ontologies and the database and to execute SPARQL queries.

For the patents case study, the retrieval system maintains several indexes related to the metadata and the claims content of the patent documents. As shown in the interface described in Section 3.1, all the documents matching the language of the user's query and the query's search criteria are selected to be the response of the retrieval system. To do so, the documents are annotated according to the structure of the two main ontologies used

[3] http:\\exopatent.ontotext.com

[4] http:\\molto.ontotext.com

for the patent domain and then, they are linked to the RDF indexes and stored according to their language. Figure 7 and Figure 8 show the structure of both ontologies related to the patent case study. The former describes the class hierarchy of the ontology, while the latter describes the relations between the concepts of the ontology. Both ontologies can be extended according to further needs of the queries. So far, we can see that they capture several aspects of the patent including the bibliographic data, such as the applicant and the expiration date, and other concepts specific to the biomedical domain, such as the active ingredient and the route of administration.

The instances that are loaded on this ontologies are taken from the FDA[5] Orange Book[6], MeSH[7], UMLS Metathesaurus[8], SNOMED CT[9] and ICD 10[10]. All these instances are also used for semantic annotation of patents while also they are populated in gazetteers. These gazetteers are used in the GATE[11] pipelines for the patent annotation task.

In the preliminary version of the prototype, the RDF indexes contain a small set of patent documents having English and French content, either original or translated. To ease the access to the content of the database, we provide here few examples of titles and description excerpts of the patent documents included in the databases:

Artificial blood and other gas transport agents[12]. This invention relates to aqueous compositions containing perfluorocyclocarbons having particular utility as artificial blood and other gas transport agents.

Preparation of functional human urokinase polypeptides[13]. The present invention relates to human urokinase polypeptide, to novel forms and compositions thereof and particularly to means and methods for the preparation in vitro of functional polypeptide species of human urokinase.

Method of making a substrate comprising a coloured composition by using an improved mutable composition and so produced substrate[14]. The present invention relates to a coloured composition. In some embodiments, the coloured composition may be employed in an electrophotographic toner, e.g., a toner employed in a photocopier which is based on transfer xerography.

Prevention of immune related removal of cells from the mammalian body, mutant PML molecules useful therefor[15]. The invention provides means and methods for at

[5] Food and Drugs Administration

[6] http://www.fda.gov/Drugs/InformationOnDrugs/ucm135821.htm

[7] http://www.nlm.nih.gov/pubs/factsheets/mesh.html

[8] http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/index.html

[9] http://www.nlm.nih.gov/research/umls/licensedcontent/snomedctfiles.html

[10] http://apps.who.int/classifications/apps/icd/icd10online/

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

[12] Bibliographic data: EP0091820 (A1) -- 1983-10-19

[13] Bibliographic data: EP0092182 (A2) -- 1983-10-26

[14] Bibliographic data: EP1020478 (A1) -- 2000-07-19

[15] Bibliographic data: EP1020520 (A1) -- 2000-07-19

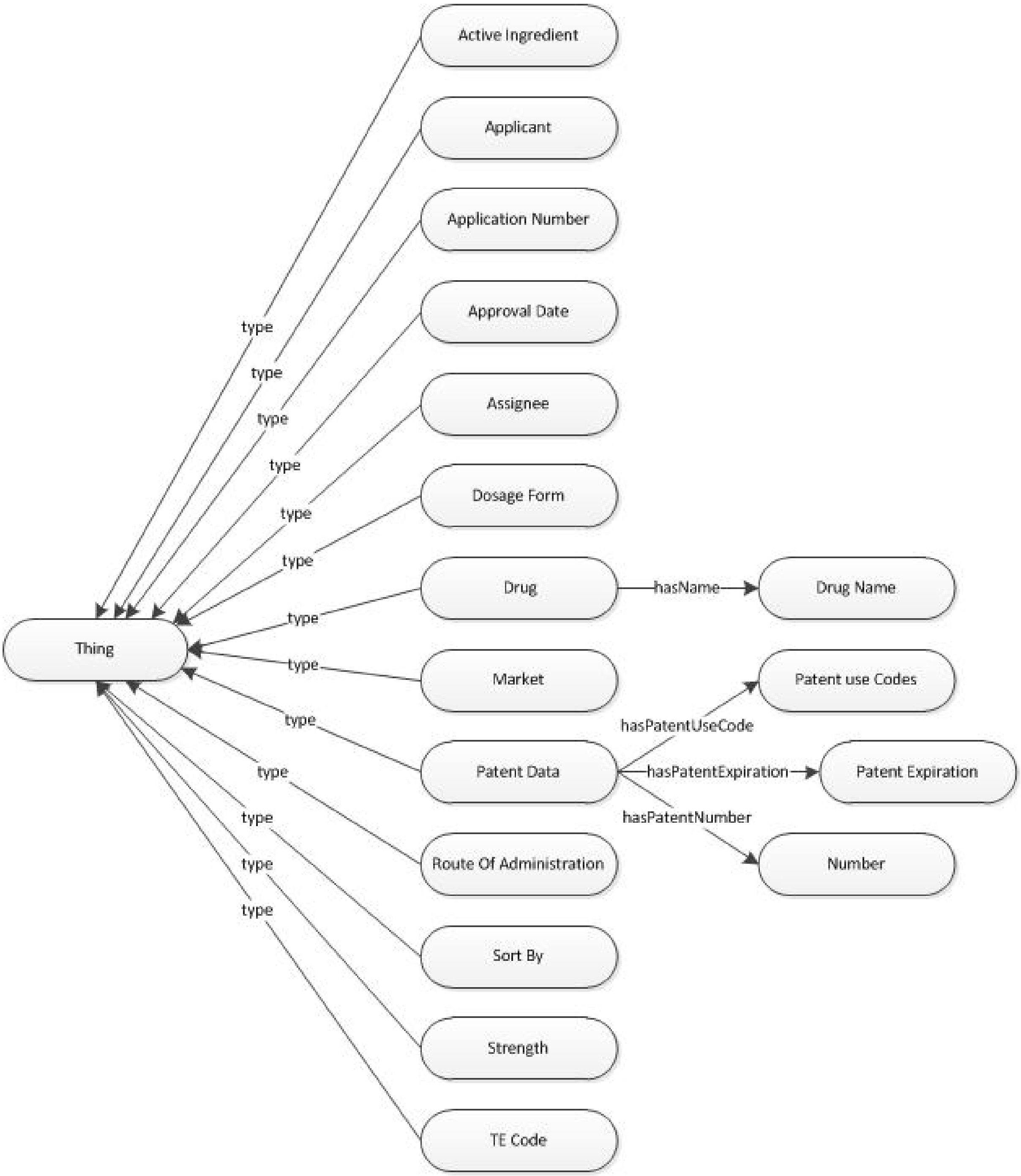


Figure 7: Class hierarchy of the ontology

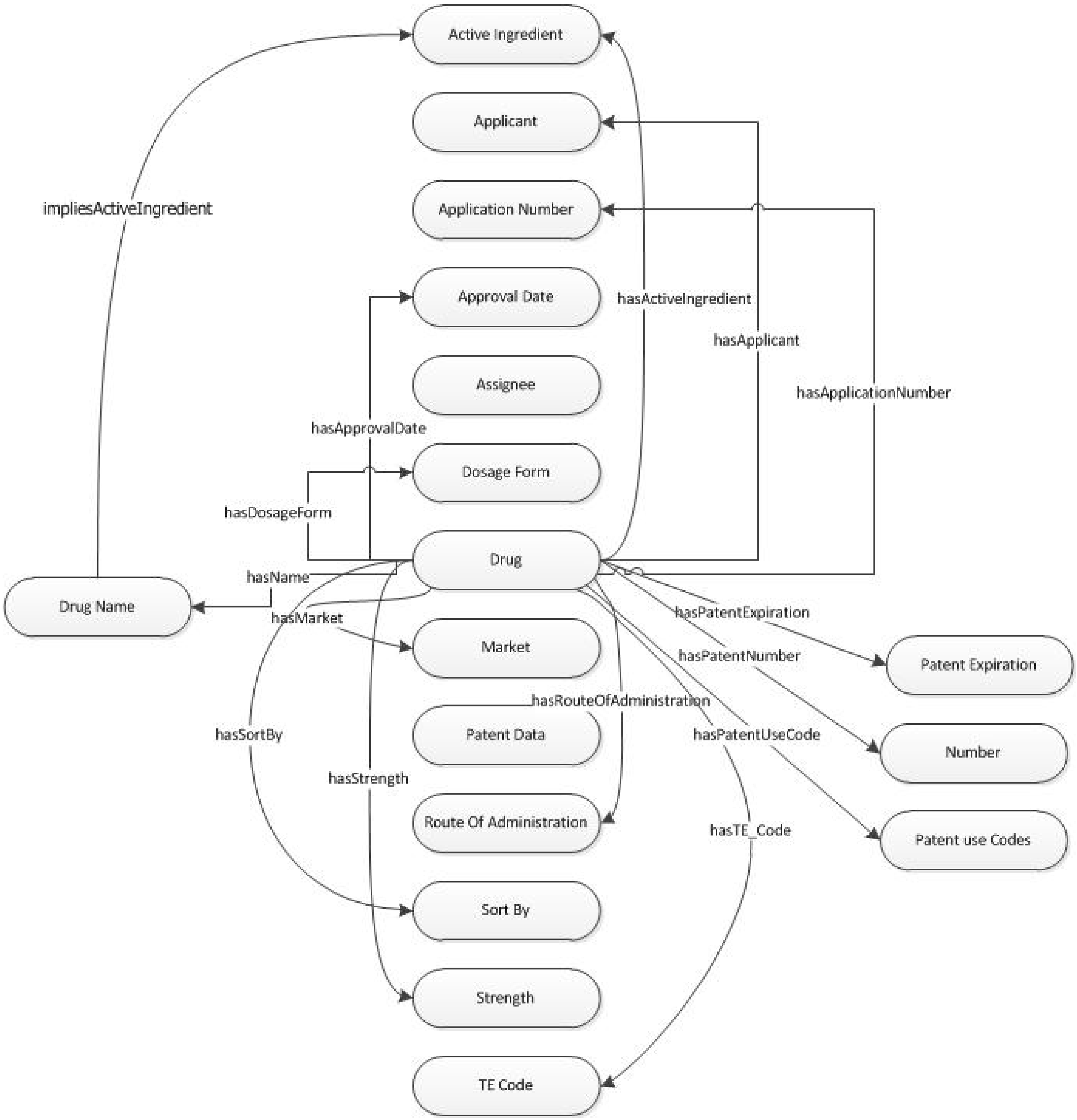


Figure 8: The patent-related concepts in the ontology

least in part preventing an immune response to certain cells in a body while leaving the general capacity of the immune system to respond to other antigens and cells essentially intact.

3.3 The Patent Translation System

The translation of patents is the second scenario in the case study. Patent documents are translated off-line before being included in the database.

The engine used for the translation has been developed within WP5 and more information is available in its corresponding documentation (Deliverable 5.2[MOL12]). The current version of the prototype uses a phrase-based statistical machine translation (SMT) system adapted to and trained on the selected domain.

The SMT system has been built using standard freely available software. A 5-gram language model is estimated using interpolated Kneser-Ney discounting with SRILM [Sto02]. Word alignment is done with GIZA++ [ON03] and both phrase extraction and decoding are done with the Moses package [KSF$^+$06, KHM$^+$07]. The optimisation of the weights of the model is trained with MERT [Och03] against the BLEU [PRWZ02] evaluation metric.

In order to adapt the system, it has been trained on parallel patents in the biomedical domain, those with IPC code A61P (see Deliverable 5.1 for more references about the corpus). Also, a preprocessing for dealing with compounds and a specialised tokenisation has been applied [ES11]. The resulting system has been evaluated using a collection of lexical metrics and showed a clear improvement with respect to the performance of non-specialised state-of-the-art SMT systems [EES$^+$11].

4 Summary and Future Directions

This document presents the preliminary prototype for the patents case study. The initial tasks include the definition of the architecture for the prototype and the two basic use case scenarios: the multilingual retrieval of biomedical patents and the translation of patent claims and abstracts.

The online demo allows several search options including natural language queries. The initial set of allowed queries include 21 different topics in relation to the biomedical domain. The grammar developed to process the queries covers about 600 queries in English and 500 in French.

In relation to the multilingual retrieval system, we describe the ontology used to deal with the biomedical domain and the extraction of FDA terms, drugs and measurement related models, and the ontology created to capture the structure of patent documents. Patents in the retrieval engine are annotated following the two main ontologies selected for the domain, besides to the general PROTON ontology. The architecture of the multilingual patents retrieval system is based on Exopatent, a working KRI platform built by Ontotext.

Regarding the SMT system, we have been working with provisional data in order to create the first version of GF grammars and the baseline of the SMT system. The recent work related to this task is currently under study in WP5, which will provide a hybrid translation system for the final prototype.

Future directions, involving all the participants of WP7, includes the development of the resources for German, the integration of the MT system into a pipeline and the design of an online process for future development, the study of the interoperability between the two technologies towards an integrated online multilingual patent retrieval and the evaluation of the systems and the resources generated.

Moreover, Ontotext will extend the current annotation tools and to adopt the Semantic Biomedical Tagger[16]. This change will involve changes in the main ontologies that are currently used, which will bring the possibility of extending the types of queries that can be expressed in natural language. An additional functionality to be added is the free text search and the combination of free text search and natural language queries. This will help the user to search for things that cannot be expressed using the current controlled language covered by the GF.

References

[EES$^+$11] Cristina España-Bonet, Ramona Enache, Adam Slaski, Aarne Ranta, Lluís Màrquez, and Meritxell Gonzàlez. Patent translation within the molto project. In *Proceedings of the 4th Workshop on Patent Translation, MT Summit XIII*, pages 70--78, Xiamen, China, sep 2011.

[ES11] Ramona Enache and Adam Slaski. Towards a patents translation system — results and perspectives, May 2011.

[GM10] 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.

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

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

[16] `http://www.ontotext.com/life-sciences/semantic-biomedical-tagger`

[MOL11] MOLTO. D5.1. description of the final collection of corpora., August 2011.

[MOL12] MOLTO. D5.2. description and evaluation of the combination prototypes., March 2012.

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

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

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

[Ran11] Aarne Ranta. *Grammatical Framework: Programming with Multilingual Grammars.* CSLI Publications, Stanford, 2011. ISBN-10: 1-57586-626-9 (Paper), 1-57586-627-7 (Cloth).

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