MOLTO

WP7: Patents Case Study

Meritxell Gonzàlez Bermúdez

2nd Year Review

Barcelona, March 20th, 2012

Objectives

- To create a prototype of MT and NL retrieval of patents
  - in the bio-medical & pharmaceutical domains,
  - allowing translation of patent abstracts & claims in English, French and German,
  - exposing several cross-language retrieval paradigms on top of them.

Workplan

No.	Title	Date
D7.1 Prototype	**Patent MT and Retrieval Prototype Beta**	**M21**
D7.2 Prototype	Patent MT and Retrieval Prototype	M27
D7.3 Report	Patent Case Study Final Report	M33
M9	Case Study Complete	M33

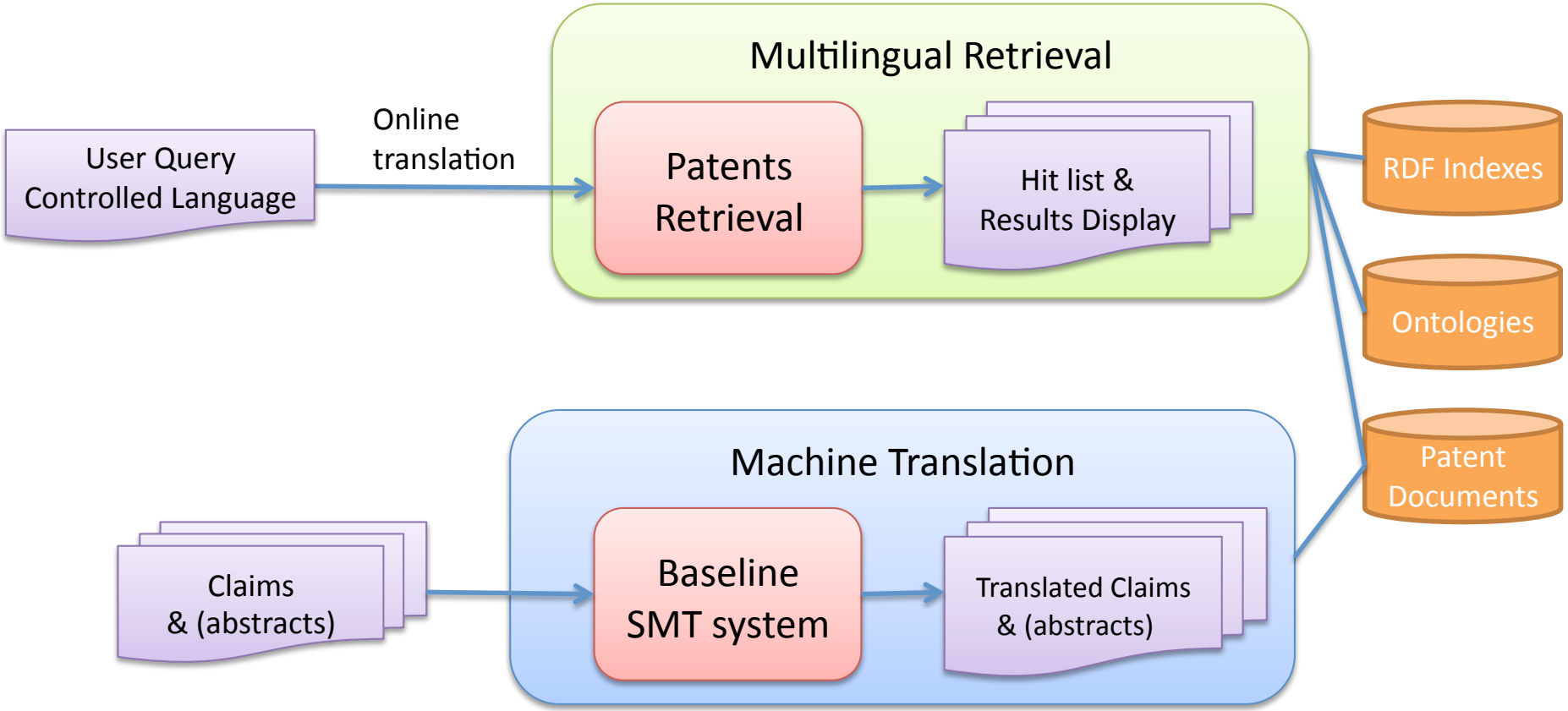
Participants

Partners	PM	Tasks
UPC	15	• Corpus building • Patents translation • MT Automatic Evaluation
Ontotext	15	• Semantic Infrastructure • Patents annotation & indexing • Prototype building
UGOT	12	• Domain Grammar

Tasks

TASK	Name
7.1	User Requirements
7.2	Corpora
7.3	Grammars for the patent domain
7.4	Ontology and Document Indexation
7.5	Patents Retrieval System
7.6	Machine Translation Systems
7.7	Prototype building (Online User Interface)
7.8	Evaluation

T7.1 - Use Case Scenarios



T7.2 - Corpora

- Official EPO Corpora (test set)
  - 66 patents belonging to the biomedical domain.

- Corpus of 7705 document retrieved from EPO website (retrieval database)
  - 4,274 out of the 7,705 documents have claims (6M lines),
  - 2,058 out of them are trilingual (3M lines).
  - 2,116 documents have claims written only in English
  - 66 have claims only in German (260K lines)
  - 34 only in French (88K lines).

- Work in progress
  - Preparing the data for translation. Currently we have FR2EN.

T7.3 - Grammar

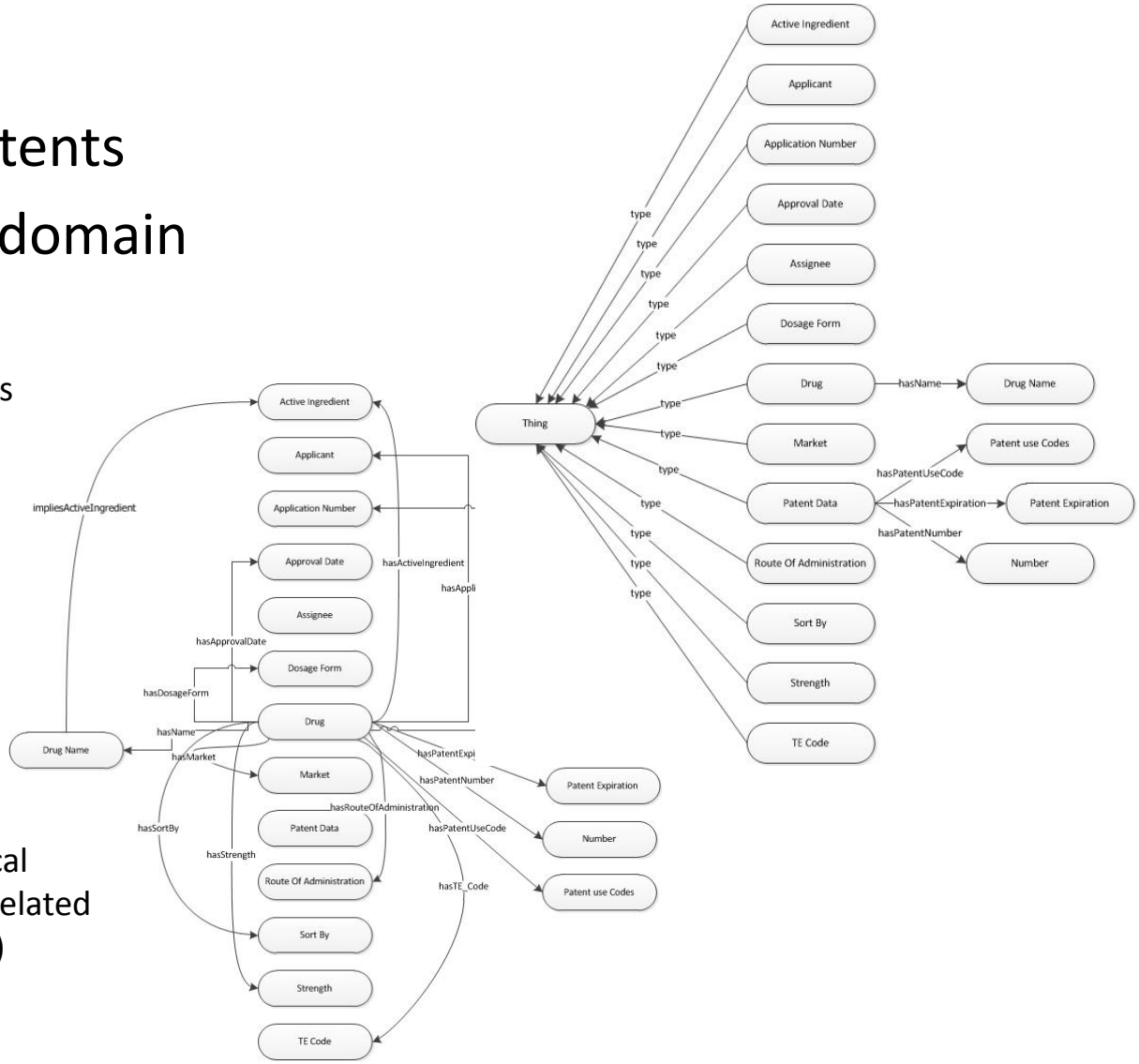
- GF grammars for Patent translation
  - Already discussed at WP5
  - Future work
    - The German version
- GF grammars for controlled language queries
  - 131 query types
  - English and French Grammars available in the beta prototype
  - Full coverage of the examples.

    ~500 sentences in French

    ~600 sentences in English
  - Future work
    - The German version

T7.4 – Ontologies

- Class hierarchy for patents
- Ontology biomedical domain
- Data models:
  - Food and Drugs Administrations Orange Book
  - MeSH (National Library of Medicine's controlled vocabulary thesaurus)
  - UMLS Metathesaurus (Unified Medical Language System)
  - SNOMED CT (Systematized Nomenclature of Medicine, Clinical Terms)
  - ICD 10th (International Statistical Classification of Diseases and Related Health Problems 10th Revision)



T7.5 –Retrieval System

- The ontologies, indexes, databases and retrieval engines have been set up for the specific domain and using bunch of patents.

- The semantic annotation process is carried by a GATE pipeline on the English texts.

- Future work:
  - Annotation of machine translated documents

T7.6 - Machine Translation

- SMT baseline system trained on the domain with the MAREC corpus:
  - FR -> EN ✓
  - DE -> EN ~
  - EN -> DE ✗
  - EN -> FR ✗

- Work in progress:
  - Improve the segmentation process

- Future work:
  - Export the semantic annotations during the translation

T7.7 – Online Demo

- Fully functional version of the prototype at http://molto-patents.ontotext.com/
- The demo allows querying the system in English and French.
- The interface allows accessing the system in three different ways:
  - the controlled language,
  - SPARQL and
  - Index terms.

T7.7 – Online Demo

- Work in progress
  - Add the new corpus to the database
  - Include the French automatic translations
  - Integrate Speech recognition
  - Extend the prediction of the controlled language
- Future work:
  - Include free text and a combination of it with the controlled language.
  - Show original text and automatic translations

T7.8 - Evaluation

- Evaluation in WP7 involves three modules:
  - Translation system
    - Human Evaluation of the translations using the TAU criteria (WP9)
    - Automatic Evaluation of the translations
  - Retrieval system
    - Automatic evaluation by means of F1 or average precision.
    - Requires manual annotation of a test set
  - The interface
    - Human evaluation of Usability or User satisfaction.
    - Requires hiring users, but we need Patent skilled users!

Dissemination

- Refereed Conferences
  - The Patents Retrieval Prototype in the MOLTO project
    Milen Chechev, Meritxell Gonzàlez, Lluís Màrquez, Cristina España-Bonet.
    Worl Wide Web Conference 2012
    16th-20th April 2012, Lyon, France

  - Patent Translation within the MOLTO project,
    Cristina España-Bonet, Ramona Enache, Adam Slasky, Aarne Ranta, Lluís Màrquez & Meritxell Gonzalez,
    MT Summit XIII - 4th Workshop on Patent Translation.
    September 23, 2011 Xiamen, China

MOLTO

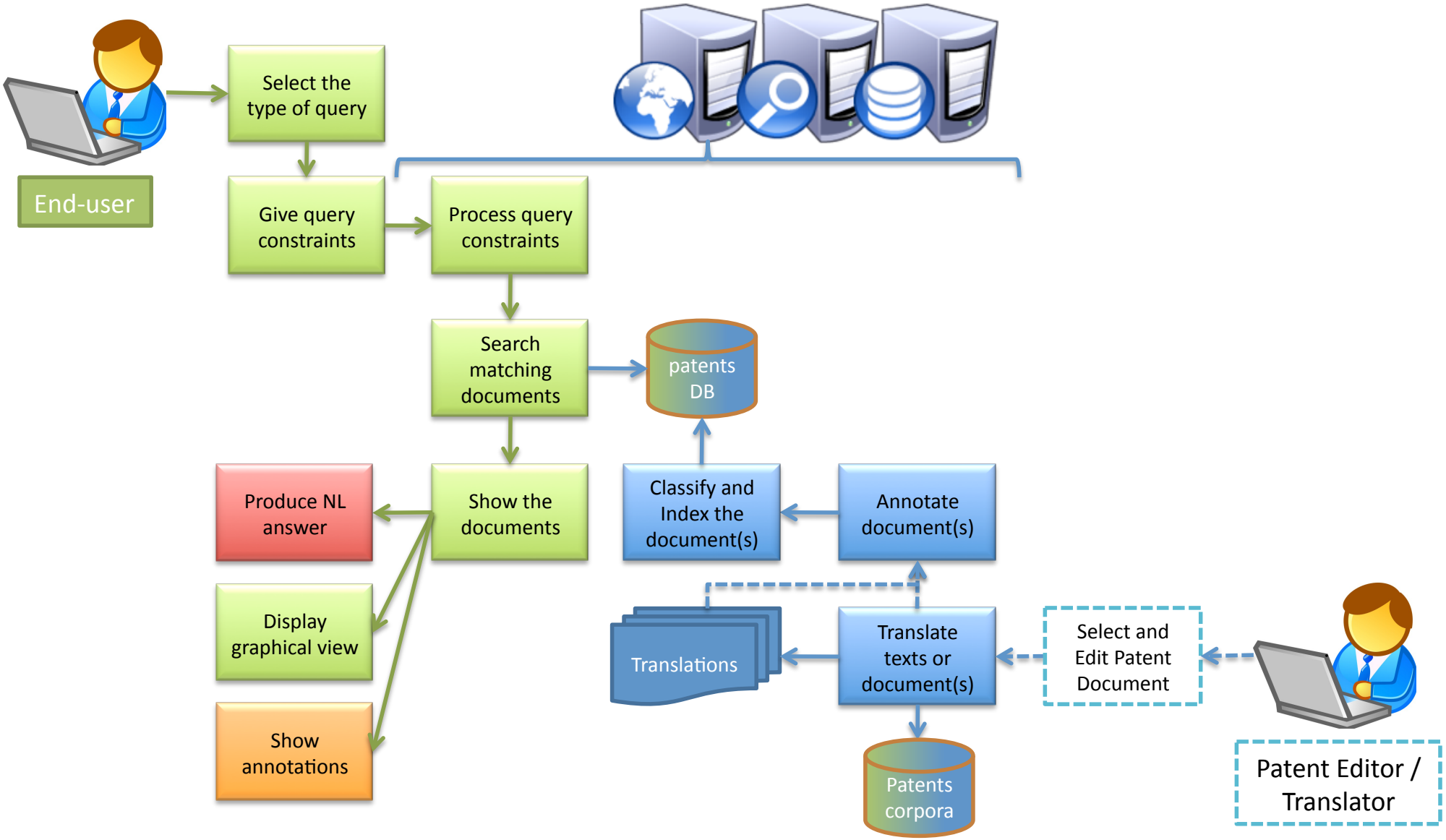
WP7: Patents Case Study

Meritxell Gonzàlez Bermúdez

2nd Year Review

Barcelona, March 20th, 2012

T7.1 - Basic Flow



T7.3 - NL Generation

- We defined the need for generating a simple NL response in the interface.

- To do so, the work to be done includes the generation of templates for each topic and the specific grammar.

Queries Examples (131 sentences)

what information can I get about A_DRUG (aspirin)
what chemical substances there are in A_DRUG?
what are the active ingredients of A_DRUG (aspirin)
give me the drugs that are compounds
what are the dosage forms of A_DRUG (aspirin)
the drug preparations for A_DRUG with a patent that expires after DATE
what is the route of administration of A_DRUG (aspirin)
I want the name of A_DRUG with a patent with approval date DATE
what is the dosage form of A_DRUG (aspirin)
what methods are used in THE_PATENT?
what is the patent number of the patent for A_DRUG
give me the use of patents approved in DATE / on DATE / before DATE / after DATE
when does THE_PATENT expire?
give me the use codes of THE_PATENT