MOLTO

WP7: Patents Case Study

Meritxell Gonzàlez Bermúdez

4th Project Meeting
Zürich, March 8th, 2012

Objectives

- To create a prototype of MT and 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.

PM

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

Deliverables

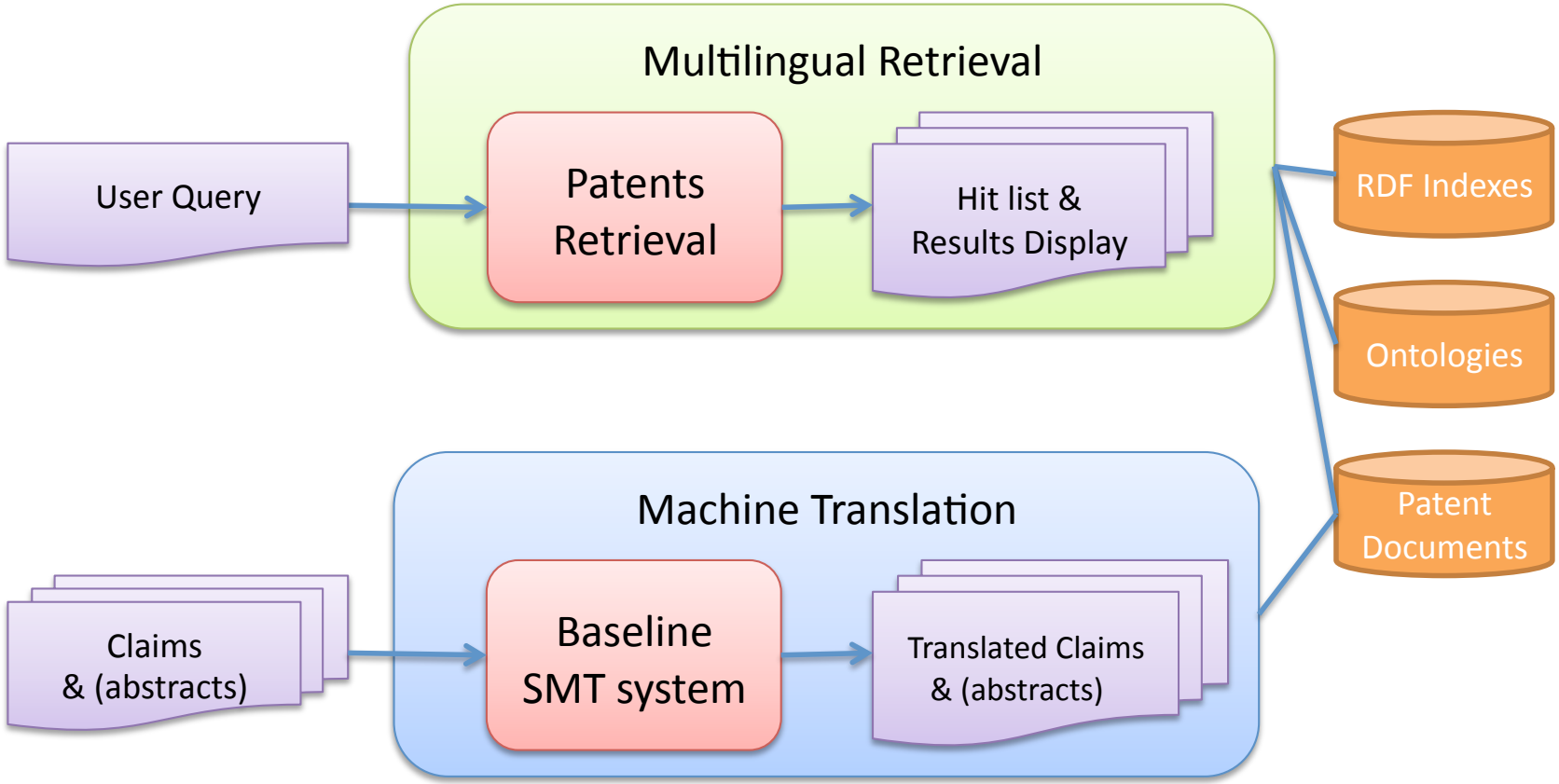
No.	Title	Date
D7.1 Prototype	**Patent MT and Retrieval Prototype Beta**	**M21**
D7.2 Prototype	**Patent MT and Retrieval Prototype**	**M27**
D7.3 Report	Patent Case Study Final Report	M33

Tasks' Progress

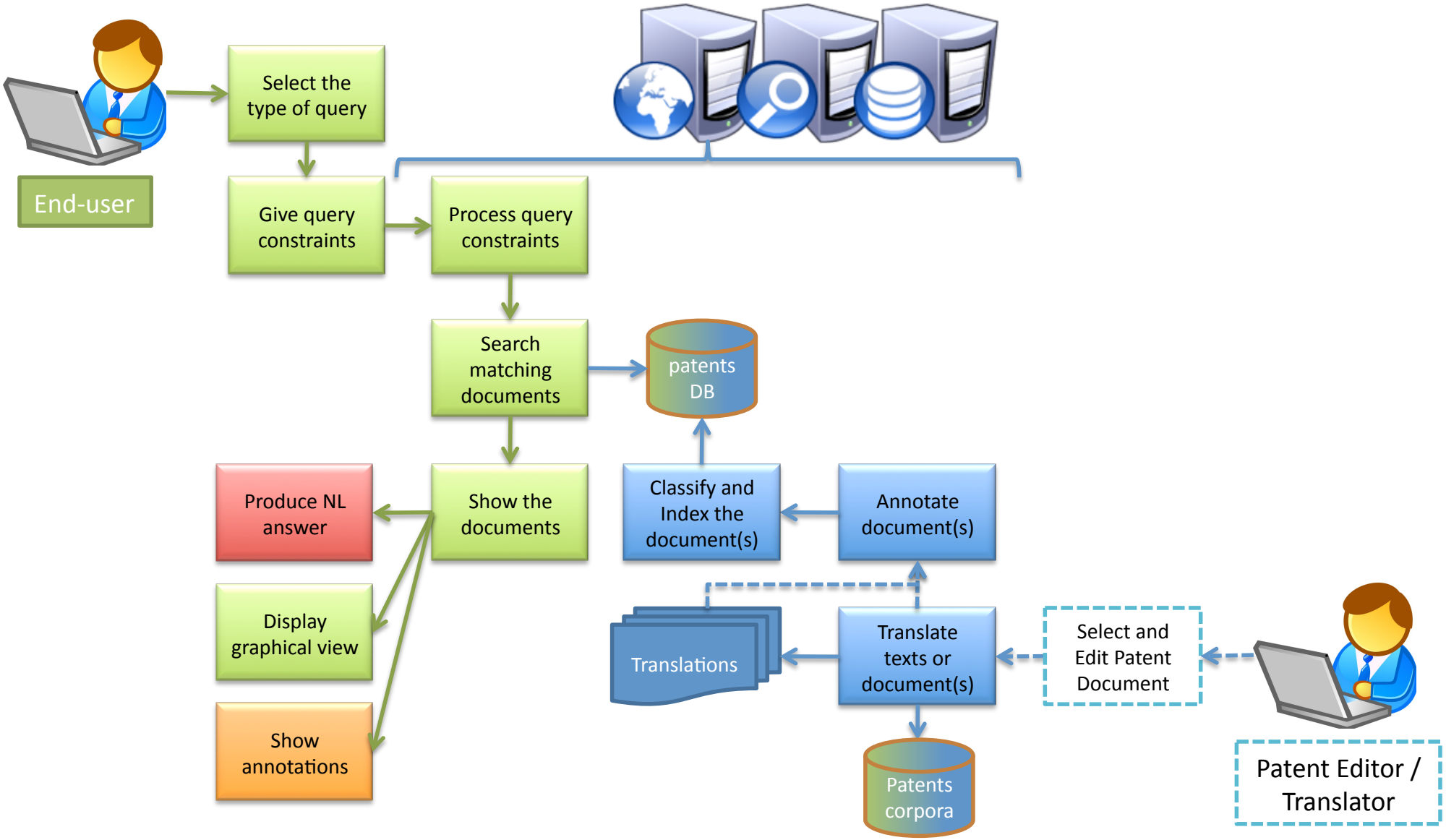
Tasks

TASK	Name	Leader	Participants
7.1	User Requirements	All	WP9
7.2	Corpora	UPC	WP5
7.3	Grammars for the patent domain	UGOT	
7.4	Ontology and Document Indexation	Ontotext	UGOT
7.5	Patents Retrieval System	Ontotext	
7.6	Machine Translation Systems	UPC	UGOT, WP5
7.7	Prototype (User Interface)	Ontotext	UGOT & UPC
7.8	Evaluation	EHU	All
7.9	Reports	All	

T7.1 - Use Case Scenarios



T7.1 - Basic Flow



T7.2 - Corpora

- Official EPO Corpora.
  - 66 patents belonging to the biomedical domain.
  - Proposal: Use this as our test set.

T7.2 - Corpora

- Alternative corpus of 7705 document directly from their website (i.e. publicly available).
  - 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).
  - Proposal:
    - Use this for building a new Language Model.
    - Include all these in the retrieval system
  - Work in progress
    - Preparing the data for translation. Currently we have FR2EN.

T7.3 - Grammar

- Patent text grammars
  - Already discussed at WP5
- Queries 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

T7.3 - Query Language

- English and French Grammars available in the beta prototype
  - Full coverage of the examples.
    - ~500 sentences in French
    - ~600 sentences in English

- The German version will be also addressed in the next months.

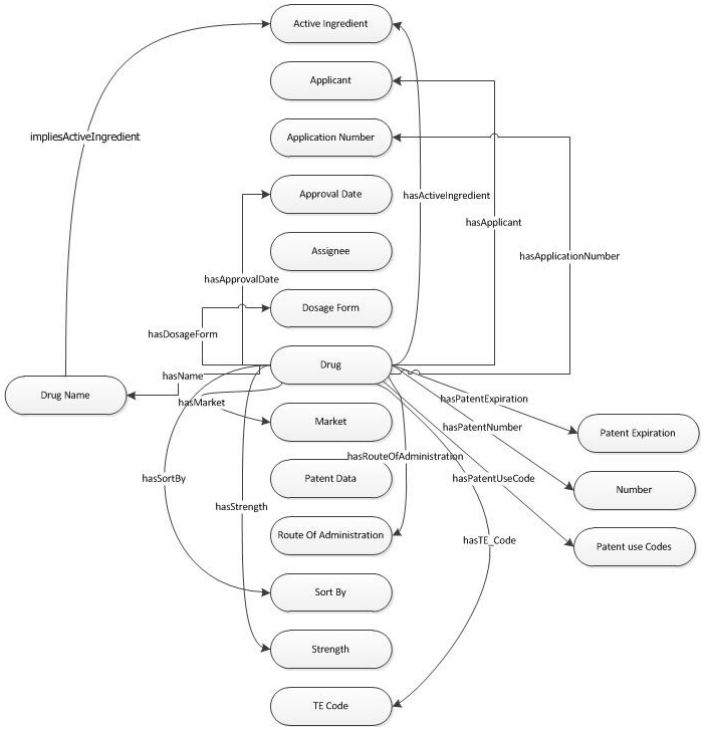
T7.3 - NL Generation

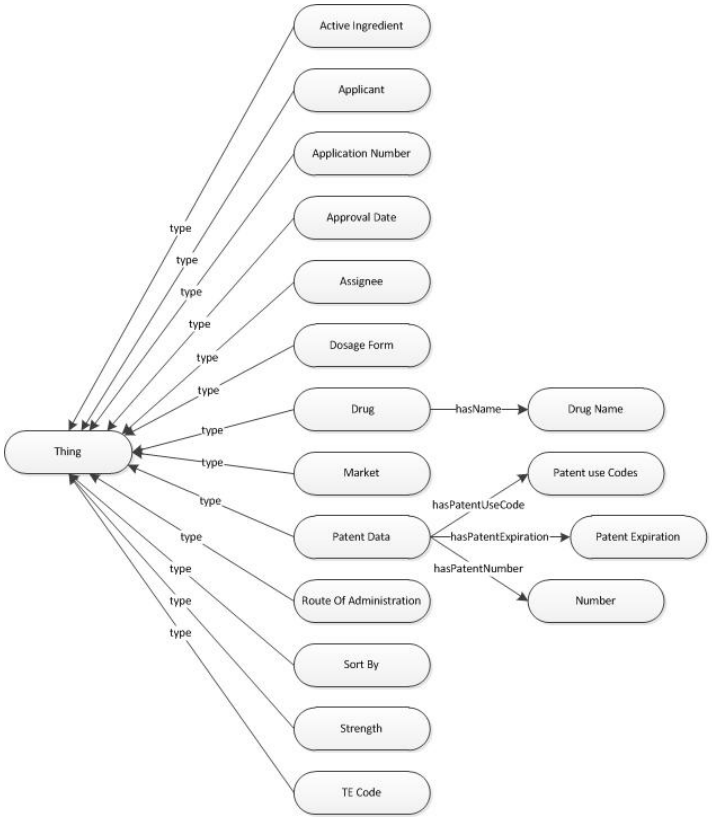
- We defined the need for generating a simple NL response in the interface.

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

T7.4 – Ontologies

- Class hierarchy
- Relations between concepts





T7.4 – Ontologies

- Food and Drugs Administrations Orange Book
  - http://www.fda.gov/Drugs/InformationOnDrugs/ucm135821.htm
- MeSH
  - MeSH is the National Library of Medicine's controlled vocabulary thesaurus in a hierarchical.
  - http://www.nlm.nih.gov/pubs/factsheets/mesh.html
- UMLS Metathesaurus
  - Unified Medical Language System
  - http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/index.html
- SNOMED CT
  - Systematized Nomenclature of Medicine--Clinical Terms
  - http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html
- ICD 10th
  - International Statistical Classification of Diseases and Related Health Problems 10th Revision
  - http://apps.who.int/classifications/icd10/browse/2010/en

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.

T7.6 - Machine Translation

- SMT hybridization is being carried in WP5 tasks.
- In WP7, we are using the SMT baseline system trained on the domain with the MAREC corpus
  - FR -> EN ✓
  - DE -> EN ~
  - EN -> DE ✗
  - EN -> FR ✗
- Work in progress:
  - Segmentation of the patent text for translation and for saving the XML structure.
  - Exportation of the semantic annotations during the translation process in order to be able to show the annotations also in the French and German texts.
  - Also, semantically aimed automatic 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 patents in the database has original text in English, French and German.
- Soon, It will include the French automatic translations!

T7.7 - Interface

- The interface allows accessing the system in three different ways:
  - the controlled language,
  - SPARQL and
  - Index terms.

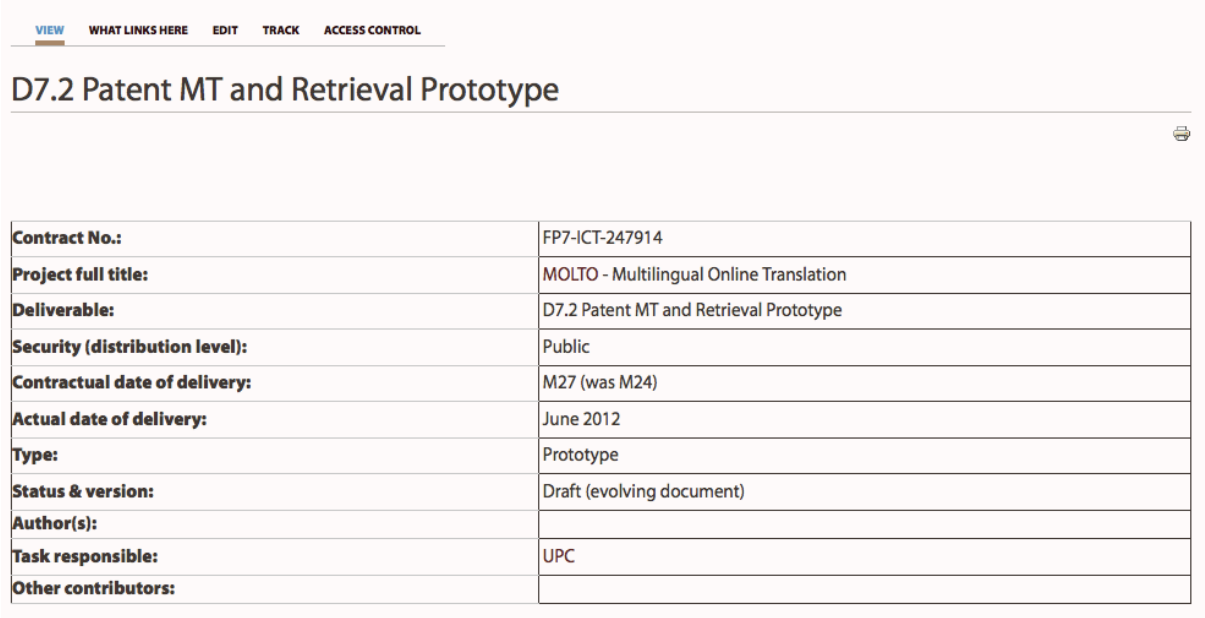
- Work in progress:
  - Speech recognition
  - Extend the prediction of the controlled language
- Future work:
  - Include free text and a combination of it with the controlled language.
  - Show the NL responses.
  - Show original text and automatic translations

T7.8 - Evaluation

- Evaluation in WP7, towards the study of the feasibility of the prototype as a part of a commercial patent retrieval system, involves three modules:
  - Translation system
    - Human Evaluation of the translations using the TAU criteria
      - Includes Hiring translators
      - Producing guidelines for translators
    - Automatic Evaluation of the translations
      - Using the selected test set
      - Running a tool like Asiya (formerly IQmt)
  - 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!

T7.9 - Reporting

- **D7.1 Patent MT and Retrieval Prototype Beta**,
  Milen Chechev, Ramona Enache, Cristina España-Bonet, Meritxell Gonzàlez, Lluís Màrquez, Borislav Popov, Aarne Ranta
  January 2012.
- **D7.2 Patent MT and Retrieval Prototype**
  http://www.molto-project.eu/wiki/living-deliverables/d72-patent-mt-and-retrieval-prototype



VIEW WHAT LINKS HERE EDIT TRACK ACCESS CONTROL

D7.2 Patent MT and Retrieval Prototype

Contract No.:	FP7-ICT-247914
Project full title:	MOLTO - Multilingual Online Translation
Deliverable:	D7.2 Patent MT and Retrieval Prototype
Security (distribution level):	Public
Contractual date of delivery:	M27 (was M24)
Actual date of delivery:	June 2012
Type:	Prototype
Status & version:	Draft (evolving document)
Author(s):	
Task responsible:	UPC
Other contributors:	

Planning

D7.2 due M27

Patent MT and Retrieval Prototype

Calendar

November (M21)	December (M22)	January (M23)	February (M24)
	D7.1. Beta Prototype T7.2. Official Parallel EPO Corpus		T7.2. www-epo extracted documents
March (M25)	**April (M26)**	**May (M27)**	**June (M28)**
T7.6 Translation of the corpora	T7.5 &T7.6 Annotation export	T7.3 German grammars T7.8 Modules evaluation **D7.2. Final Prototype**	T7.8 Evaluation starts
July (M29)	**August (M30)**	**September (M31)**	**October (M32)**
	D73. Final Report		

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

AOB