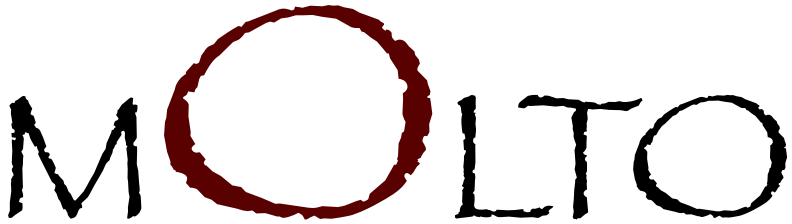


WP7: Patents Case Study

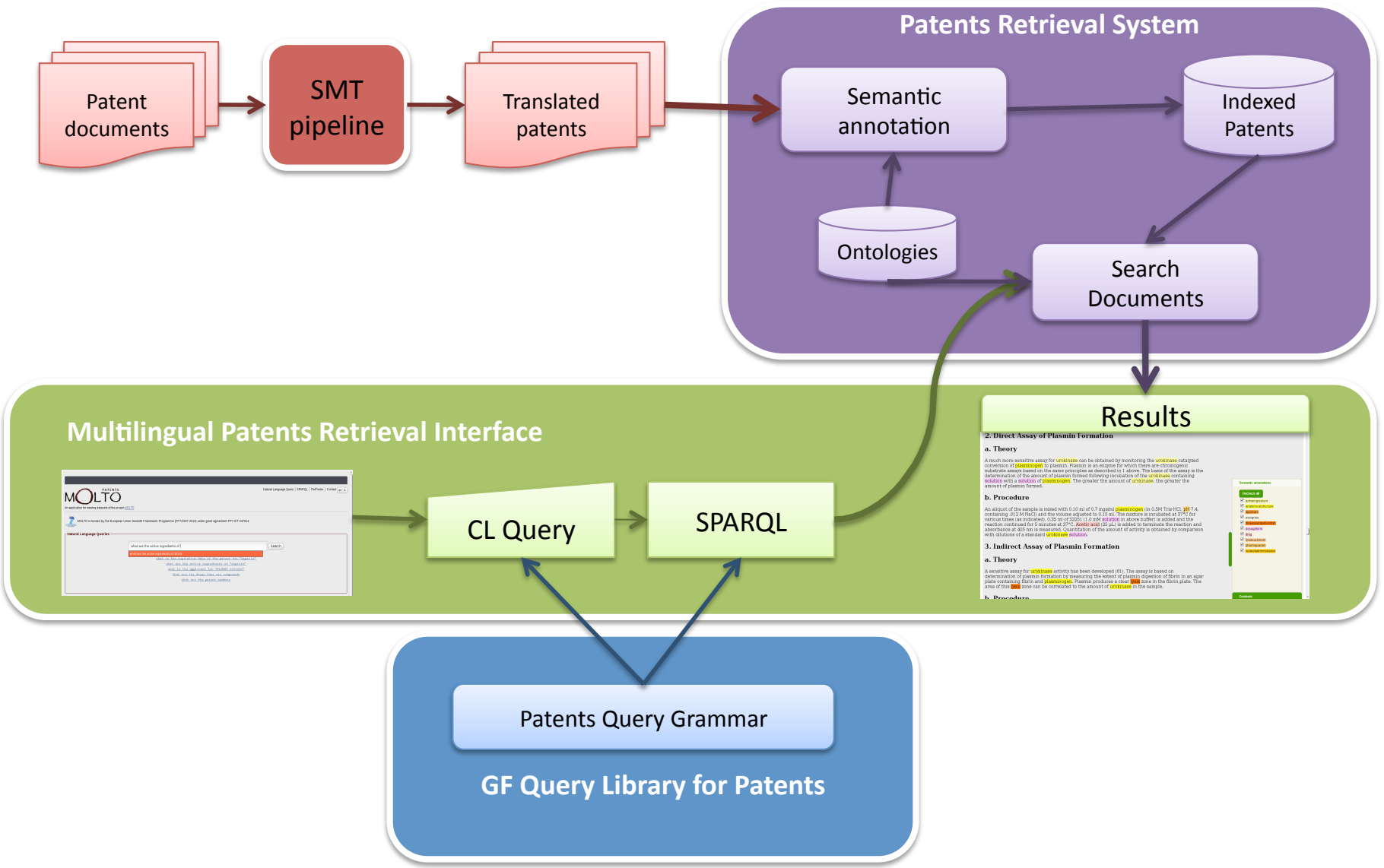
Meritxell Gonzàlez Bermúdez

5th Project Meeting
Utrecht, September 20th, 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.

Prototype architecture



Tasks' Progress

Tasks

TASK	Name	Completion
7.1	User Requirements	100%
7.2	**Corpora**	**85%**
7.3	**(query) Grammars for the patent domain**	**80%**
7.4	Ontology and Document Indexation	100%
7.5	Patents Retrieval System	100%
7.6	**Machine Translation Systems**	**90%**
7.7	**Prototype (User Interface)**	**99%**
7.8	**Evaluation**	**0%**

T7.2 - Corpora

- The patents downloaded from the EPO website (~7705 files) have been automatically translated using the baseline SMT system and semantically annotated.

- The complete collection of files consists of
  1. the original patent documents,
  2. the English version of the patent documents having the semantic annotations, and
  3. the automatic translations of claims, abstracts and descriptions.

- These documents constitute the main content of the retrieval databases.

T7.2 - Corpora

Task Completion	85%

- **Further work:**
  - **Translation of the documents keeping the semantic annotations (ongoing)**
  - **Evaluate the quality of the annotation's traslation**

T7.3 – Query Grammar

- **Y2 Review, Recommendation 7**: *WP7 work should focus on the major issues examined in MOLTO, especially in relation to the grammar – ontology interoperability automation.*

- Query grammars have been refactored using the set of primitives defined in the Query Library work conducted in WP4.
  - English and French version of the patents query grammar were adapted to the new structure
  - German version has been developed from scratch

T7.3 – Query Grammar

- In comparison to the previous patent query grammar, now it has fewer constructions, because of the fact that it is developed on top of the Query Library.

- The constructions are more natural and the number of malformed constructions have decreased considerably.

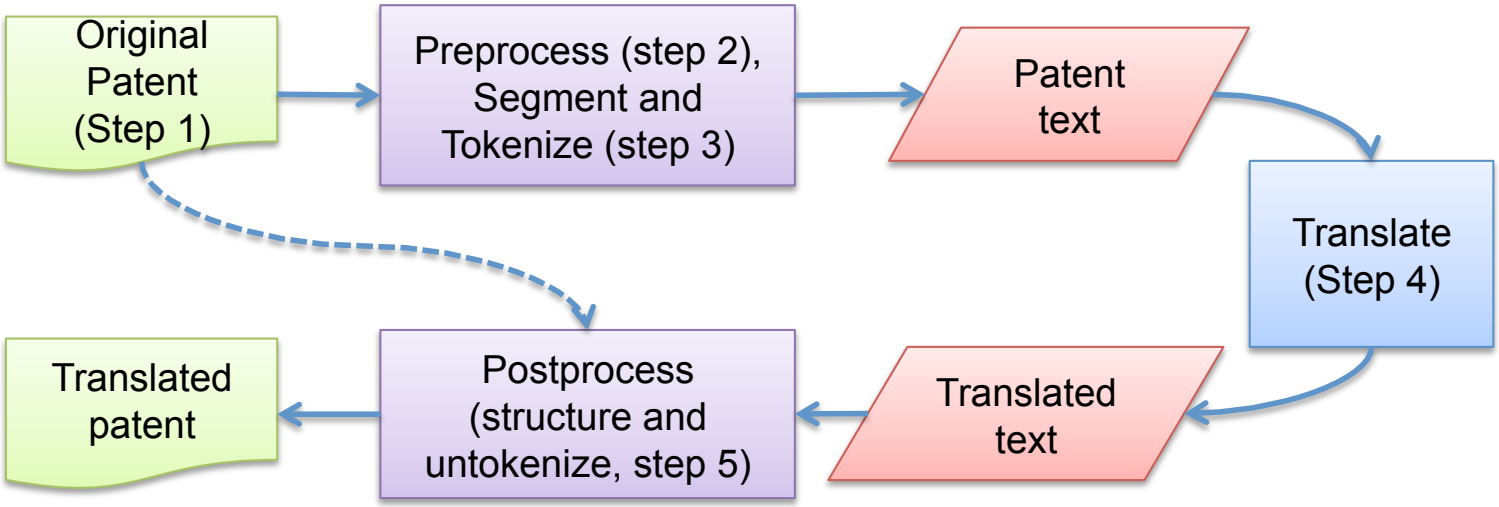
- The current grammar is able to parse/generate 359 query constructions in English, 111 in French and 147 in German.

Task Completion	80%

- **Volunteers needed to review the sentences**

T7.6 - Machine Translation

- Patent translation pipeline that allows for creating a translated document having the same XML structure as the original patent.



T7.6 - Machine Translation

- **2Y Review. Comments about the objectives:**
  - *The goal of transferring semantic annotations to the target language is unclear.*
  - *The semantic annotations could also be exploited by MT and not only for retrieval.*

Task Completion	90%

- The work in progress consists of
  1. training the SMT system with the semantic annotations on the training set
  2. translate the retrieval dataset using this new system in order to translate also the semantic annotations.

T7.7 - Interface

- Fully functional version of the prototype at http://molto-patents.ontotext.com/

- The demo allows querying the system in English, French and **German**.

- The patents in the database has original text in English and **automatic translations to French and German.**

T7.7 - Interface

- Some basic tests have been carried out in order to improve usability, resulting in several corrective actions:
  - The query examples in the demo website
  - The language swap in the document view interface
  - The review of the queries interpretation (e.g. the case of the active ingredients of *ampicilline*).

- Database roadmap. The deliverable contains a summary of the patents database in order to provide users with a comprehensive set of examples.

Task Completion	99%

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
    - Automatic Evaluation of the translations
    - Comparison between the semantic and non-semantic translations
  - Retrieval system
    - Automatic evaluation by means of F1 or average precision.
    - **Requires manual annotation of a test set**
  - The prototype (WP9)
    - Human evaluation of Usability or User satisfaction.

Task Completion	**0%**

Planning

D7.3 due M33

Patent Case Study. Final Report

Deliverables

No.	Title	Date
D7.1 Prototype	Patent MT and Retrieval Prototype Beta	M21
D7.2 Prototype	Patent MT and Retrieval Prototype	M27 (M30)
D7.3 Report	Patent Case Study Final Report	M33 (M36?)

Task Completion	66%

Calendar

March (M25)	April (M26)	May (M27)	June (M28)
	T7.5 &T7.6 Annotation export of the retrieval dataset	**D7.2. Final Prototype**	
July (M29)	**August (M30)**	**September (M31)**	**October (M32)**
T7.3 German grammars T7.6 Translation of the corpora	**D7.2. Final Prototype**	T.2 & T.6 Training and translation using the baseline SMT with semantics	T7.8 **Retrieval & annotation evaluation**
November (M33)	**December (M34)**	**January (M35)**	**February (M36)**
D73. Final Report	**D53. Final Hybrid version**	T.6 & T.8 Translation and evaluation with semantics using the **hybrid**	**D73. Final Report**

Dissemination

- **Demo presentation**
  - **CNLs for multilingual queries in MOLTO**
    Olga Caprotti, Milen Chechev, **Ramona Enache,** Meritxell Gonzalez, Aarne Ranta, Jordi Saludes
    Third Workshop on Controlled Natural Language (CNL 2012)
    29–31 August 2012, Zurich, Switzerland

- **Reports**
  - **D7.2 Patent MT and Retrieval Prototype**
    Meritxell Gonzàlez, Milen Chechev, Mariana Damova, Ramona Enache, Cristina España-Bonet, Lluis Marquez, Maria Mateva, Aarne Ranta, Laura Tolosi
    September 2012
  - **MOLTO-Patents: recent issues, solutions and perspectives**
    Laura Tolosi, Maria Mateva
    12 September 2012, Ontotext AD, Bulgaria

MOLTO

AOB

Meritxell Gonzàlez Bermúdez

5th Project Meeting

Utrecht, September 20th, 2012