WP4 Knowledge Engineering

Borislav Popov and Petar Mitankin, Ontotext

Luxembourg, 15 March 2011

WP General view
Objectives

- interoperability:

  NL $\leftrightarrow$ Formal Knwoledge

- support of MT tasks with background knowledge:

  Knowledge Representation Infrastructure

  data sets, ontologies, alignment

WP General view
Partners and Deliverables

- Ontotext: 30 PM
- UHEL: 12 PM
- UGOT: 4 PM

D 4.1 Knowledge Representation Infrastructure M8

D 4.2 Data Models, Alignment Methodology, Tools and Documentation M14

D 4.3 Grammar-Ontology Interoperability M18

WP General view
Requirements for the knowledge modeling infrastructure

- Building the conceptual models and knowledge bases needed for grammar development and the use cases of MOLTO - one base set and three specialized knowledge sets for the use cases;
- The specialized sets will include the necessary domain specific models and instances, e.g. multi-lingual patent classification taxonomies, museum ontology and instance base, etc. Using a semantic alignment methodology paired with a set of data source transformation tools for each of the structured data sources.

Ongoing work
Modules of the infrastructure

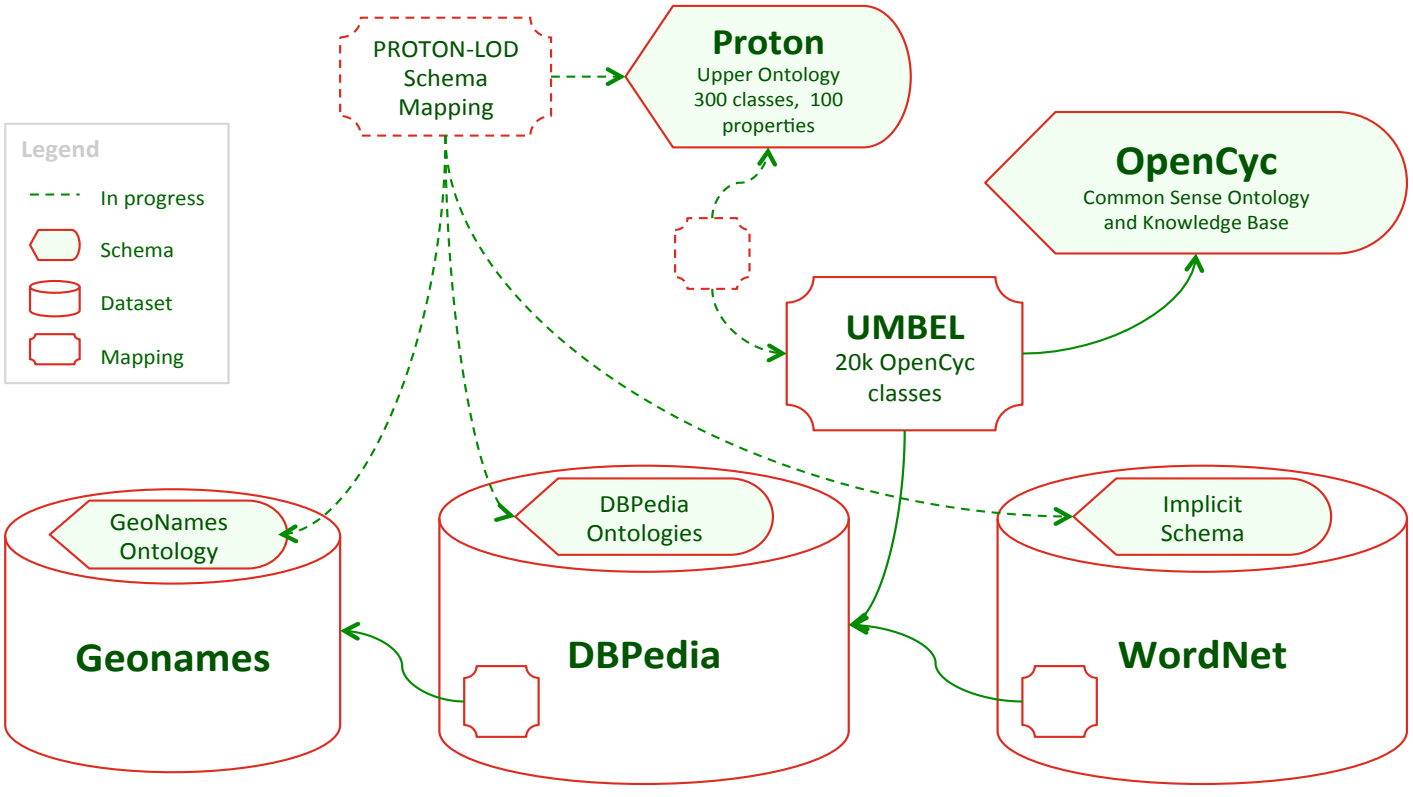
The infrastructure includes:

- OWLIM — a semantic repository that stores all structured data such as ontologies, background knowledge, etc., and provides SPARQL query mechanism and reasoning;
- RDFDB — an API that provides a remote access to the stored structured data via JMS;
- KRI Web UI — a UI that accesses OWLIM through the RDFDB layer. The web UI gives the user the possibility to browse the ontologies and the database, to execute SPARQL queries, etc.

Ongoing work
Data sets

- wkb - 29104 named entities: 6006 persons, 8259 organizations, 12219 locations and 2620 job titles
- dbpedia - 1.67 million things: 364,000 persons, 462,000 places, 99,000 music albums, ...
- wordnet, linked data, ...

Ongoing work
Reference Knowledge Stack



Ongoing work
The Natural Language-GF-Ontology Interoperability

$$\text{Natural Language} \xrightarrow{\text{GF}} \text{Trees} \xrightarrow{\text{mapping rules}} \text{Ontology}$$

$$\text{Natural Language query} \longrightarrow \text{SPARQL query to given Ontology}$$

Ongoing work
The query GF grammars

The Query Grammars provided by prof. Aarne Ranta, UGOT:

- English
- Swedish

The language represented by the Query Grammars:

give me all people
give me all organizations in *L*
give me all persons that work as *JT* at *O*
...

Ongoing work

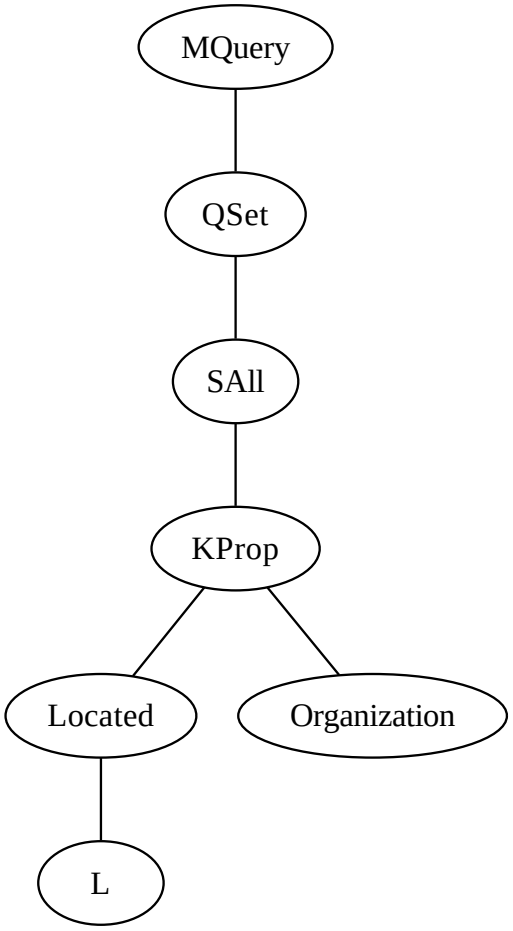
Multiple ways to say one and the same thing

64 ways to say
give me all people that work at O:

give me all persons that work at O
give me all people that collaborate in O
give me all persons that collaborate in O
give me the people that work at O
give me the persons that work at O
give me the people that collaborate in O
give me the persons that collaborate in O
give me the names of all people that work at O
give me the names of all persons that work at O
give me the names of all people that collaborate in O
give me the names of all persons that collaborate in O
give me the names of the people that work at O
give me the names of the persons that work at O

Ongoing work
GF as a parser

all organizations located in L



Ongoing work
Language for mapping rules

```
tree pattern | boolean condition -> output string

//all people, all locations, all organizations
(QSet ?X) | single(X) && type(X) == "" ->
select() sparqlVar(name(X)) WHERE  sparqlVar(name(X))
rdftype() class(name(X)) . ;

#define select() { SELECT ## "   " ## DISTINCT }

#table sparqlVar[2] {
Person ?person;
Location ?location;
Organization ?organization;
}
```

Ongoing work
Execution of the mapping rules

All mapping rules are compiled in one deterministic finite state machine.

Number of rules: 16
Number of test trees: 27
Avg time per tree: 0.37 milliseconds

Number of rules: 1956
Number of test trees: 1956
Avg time per tree: 0.25 milliseconds

Future work
Main task: automatization

corpus of queries $\overset{automatically}{\longrightarrow}$ GF grammars

corpus of queries $\overset{semi-automatically}{\longrightarrow}$ mapping rules:

- corpus of queries $\overset{automatically}{\longrightarrow}$ tree patern, boolean condition, part of output
- a person who knows the ontology *manually* complete the output part of the mapping rules

Future work
Other tasks

- Ontotext mapped DBpedia 3.6 to PROTON → adding DBpedia 3.6 to KRI
- adding new languages in the KRI
- adding TF and word-net from UHEL to KRI
- improvements in the user interface

Publications related to WP4

[Accepted, September 2010]
Enache, R., Angelov, K.: *Typeful Ontologies with Direct Multilingual Verbalization.* In 2nd Workshop on Controlled Natural Language, Marettimo Island, Sicily (ITALY) September 13-15, 2010

[Submitted, October 2010]
Enache, R., Angelov, K.: *Typeful Ontologies with Direct Multilingual Verbalization.* In Special Issue of the Studia Logica Journal on Logic and Natural Language, 2011