MOLTO

A company of Sirma Group
www.sirma.com

ontotext

The Mechanics behind the Natural Language-GF-Ontology Interoperability. Natural Language Based Semantic Queries

Borislav Popov and Petar Mitankin, Ontotext

Second Project Meeting of MOLTO

University of Gothenburg, 9 March 2011

The goal of WP4

The objectives of WP4 are

- research and development of two-way grammar-ontology interoperability bridging the gap between natural language and formal knowledge;
- infrastructure for knowledge modeling, semantic indexing and retrieval;
- modelling and alignment of structured data sources;
- alignment of ontologies with the grammar derived models.

MOLTO

A company of Sirma Group
www.sirma.com

ontotext

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.

Modules of the infrastructure

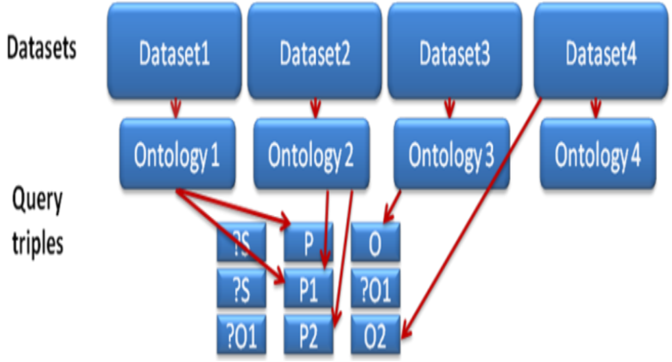
The infrastructure includes:

- 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;
- PROTON Ontology — a light-weight 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;
- 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.

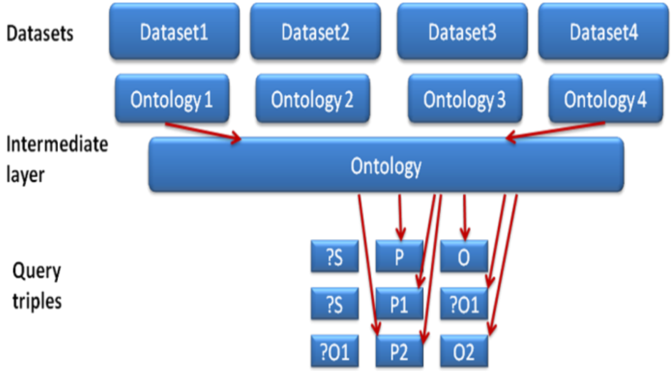
Data sets

- 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, ...
- umbel, wordnet, linked data, ...

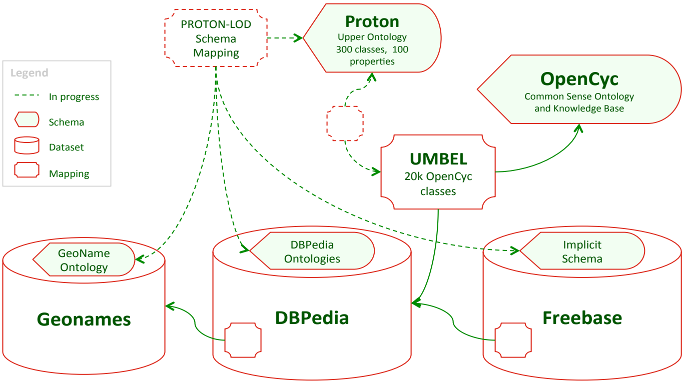
Many datasets/ontologies + no alignment



Many datasets/ontologies + alignment



The process of alignment

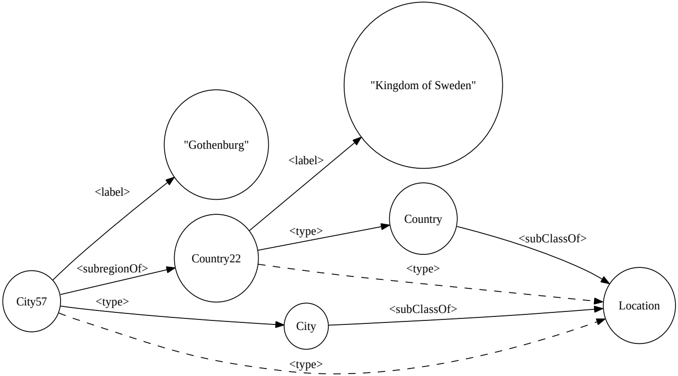


The Mechanics behind the Natural Language-GF-Ontology Interoperability

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

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

The concrete dataset + ontology as a directed graph



arcs in the graph: $500,000$

arcs + automatically inferred arcs: $1,000,000$

SPARQL

$$\frac{\text{SPARQL}}{\text{ontology}} = \frac{\text{SQL}}{\text{relational database}}$$

```
SELECT DISTINCT ?from ?label ?to WHERE {
   ?from ?label ?to .
}
```

from	label	to
$node'_1$	$label_1$	$node''_1$
$node'_2$	$label_2$	$node''_2$
...		
$node'_N$	$label_N$	$node''_N$

Example: all organizations

```
SELECT DISTINCT ?x WHERE {
   ?x <type> <Organization> .
}
```

x
$node_1$
$node_2$
$\ldots$
$node_K$

Example: all persons that work as project manager at Ontotext

```
SELECT DISTINCT ?person WHERE {
   ?person <hasPosition> ?jobPos .
   ?jobPos <withinOrganization> ?org .
   ?org <label> "Ontotext".
   ?jobPos <hasTitle> ?jobTit .
   ?jobTit <label> "Project Manager".
}
```

The query GF grammars

The Query Grammars:

15 categories: Query, Relation, Kind, Property, Individual, Activity, Name, Loc, Org, Pers, ...
59 functions: ...

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
...

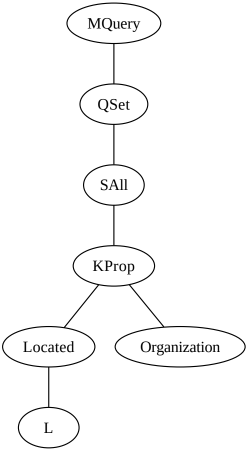
Multiple ways to say one and the same thing

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
give me the names of the people that collaborate in O

We have the grammars: GF works perfectly as a parser!

all organizations located in L



The 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;
}
```

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

- Ontotext mapped DBpedia 3.6 to PROTON. There 1.67 million things in DBpedia 3.6 that are classified in a consistent ontology, including 364, 000 persons, 462, 000 places, 99, 000 music albums, 54, 000 films, 16, 500 video games, 148, 000 organizations, 148, 000 species and 5, 200 diseases. We shall apply our MOLTO natural langauge query system to DBpedia 3.6.
- semi-automatic generation of GF grammars and mapping rules from corpus of queries
- improvements in the user interface are possible