Querying distributed cancer databases using domain concepts

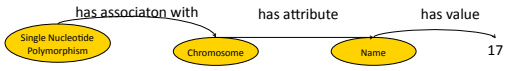
Alejandra González Beltrán[1,2], Ben Tagger[1], Anthony Finkelstein[1]

[1] Department of Computer Science [2] Computational and Systems Medicine
University College London, United Kingdom
{a.gonzalezbeltran, b.tagger, a.finkelstein}@cs.ucl.ac.uk

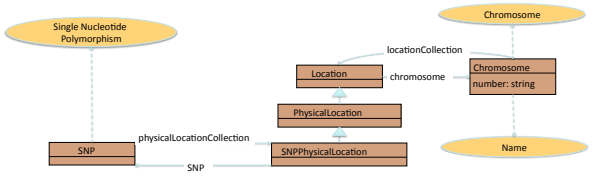
Objective: Navigational ontology-based queries on caGrid

We support high-level and descriptive queries of cancer-research data sources, based on concepts from the cancer domain such that the queries navigate the structure of the data resources without the user being aware of these structures. Our system is general but the implementation is based on the caGrid infrastructure.

For example, using the caBIO data service a user wants to retrieve Single Nucleiotide Polymorphisms associated with the Chromosome whose name is *17*.



Our system takes this query and transforms it into the corresponding query for caBIO (in CQL language) by finding the navigation paths from SNP to Chromosome in the caBIO information model.



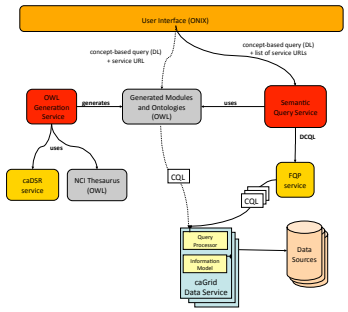
Approach

❑ **We combine Semantic Web technologies with caGrid** (US National Cancer Institute Informatics Infrastructure)

❑ For each data service, **we generate an ontology using the Web Ontology Language (OWL).** Each ontology models the metadata (this approach is also useful for data integration)
- ❑ We extract a module ontology from NCI thesaurus (NCIt), relevant for the data service
- ❑ We transform annotated UML models to OWL

❑ **We transform a query at the ontology level (a DL-query) to D/CQL**
- ❑ We rewrite and translate the query using reasoning

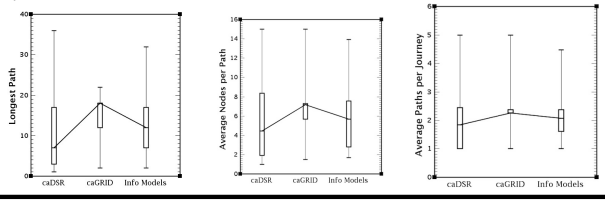
❑ We have implemented this approach using the OWLAPI and the Pellet reasoner



OWL Generator Service & Analysis of Generated Ontologies

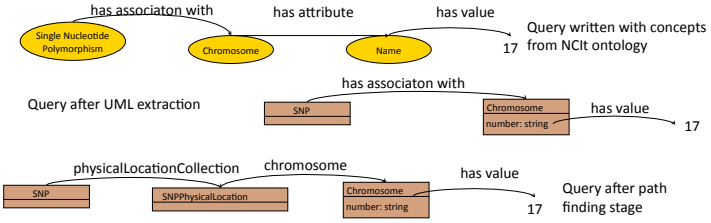
❑ We developed a caGrid Analytical Service to generate OWL ontologies from caGrid information models, as a semantic representation of the annotated UML models: OwlGen service

❑ We performed an analysis of the generated ontologies originated from: models registered in the caDSR metadata registry, models registered in the caGrid index service and models supported by deployed services (InfoModels)



Semantic Query Service

Service based on query rewriting and translation: from ontology-based to CQL queries (caGrid query language)



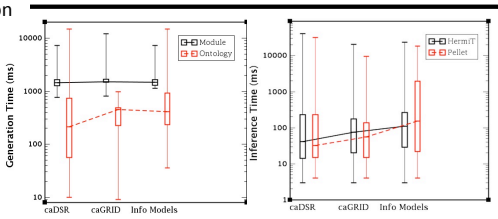
We use an intermediate calculus (Monoid Comprehension Calculus) and then translate the query to CQL.

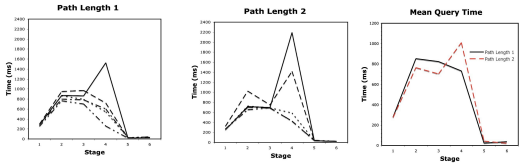
```
<ns1:CQLQuery xmlns:ns1="http://CQL.caBIG/1/gov.nih.nci.cagrid.CQLQuery">
<ns1:Target name="gov.nih.nci.cabio.domain.SNP">
<ns1:Association name="gov.nih.nci.cabio.domain.SNPPhysicalLocation"
roleName="physicalLocationCollection">
<ns1:Association name="gov.nih.nci.cabio.domain.Chromosome" roleName="chromosome">
<ns1:Attribute name="number" predicate="EQUAL_TO" value="17"/>
</ns1:Association>
</ns1:Association>
</ns1:Target>
</ns1:CQLQuery>
```

CQL query to be sent to caGrid

Performance Evaluation

Time measurements for:
-Ontologies and modules generation
- obtaining inferred ontologies using Pellet and HermiT reasoners





Query rewriting/ translation evaluation for queries of path lenghts 1, 2 and the mean query time for the two types of paths

Conclusions and Future Work

We presented our approach to support navigational ontology-based queries over caGrid (US NCI infrastructure).

We presented the **OWLGen service**, which generates OWL ontologies from caGrid UML models and analysis of path metrics over the generated ontologies. The OWL generation is parameterised and can produce **OWL2EL ontologies**. We showed the steps of the query translation process. We also presented performance evaluation results for the generation of ontologies and modules as well as their classification, and some results on query performance considering path length.

As future work, we are completing an web-based interface exposing our ontology-based querying system.

References and Acknowledgements

The authors are grateful to the National Cancer Research Institute Informatics Initiative for support for their research.

Previous work related to this approach:
[1] **"Ontology-based queries over cancer data"** SWAT4LS 2010, **Best Paper Award**, Berlin - Germany, December 8-10, 2010
[2] "**Semantic web data warehousing for caGrid**" BMC Bioinformatics, Vol 10, Supp 10, 2009.
[3]**"Domain Concept-Based Queries for Cancer Research Data Sources"**, CBMS 2009, August 3-4, Albuquerque, NM, USA
Bibliography