Ontology-based Queries over Cancer Data

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

[1]Department of Computer Science [2]Computational and Systems Medicine
University College London
London, United Kingdom

Semantic Web Applications and Tools for Life Sciences
(SWAT4LS)
Berlin, Germany
10th December 2010

Overview

- Motivation: queries over cancer data
- Background: caGrid infrastructure
  - caGrid query language (CQL)
- Objective: ontology-based queries over the caGrid infrastructure
- Approach:
  - OWL representation of caGrid models
  - Query rewriting & translation
- Implementation & performance evaluation
- Conclusions

Overview

- Motivation: queries over cancer data
- Background: caGrid infrastructure
  - caGrid query language (CQL)
- Objective: ontology-based queries over the caGrid infrastructure
- Approach:
  - OWL representation of caGrid models
  - Query rewriting & translation
- Implementation & performance evaluation
- Conclusions

Overview

- Motivation: queries over cancer data
- Background: caGrid infrastructure
  - caGrid query language (CQL)
- Objective: ontology-based queries over the caGrid infrastructure
- Approach:
  - OWL representation of caGrid models
  - Query rewriting & translation
- Implementation & performance evaluation
- Conclusions

Overview

- Motivation: queries over cancer data
- Background: caGrid infrastructure
  - caGrid query language (CQL)
- Objective: ontology-based queries over the caGrid infrastructure
- Approach:
  - OWL representation of caGrid models
  - Query rewriting & translation
- Implementation & performance evaluation
- Conclusions

Overview

- Motivation: queries over cancer data
- Background: caGrid infrastructure
  - caGrid query language (CQL)
- Objective: ontology-based queries over the caGrid infrastructure
- Approach:
  - OWL representation of caGrid models
  - Query rewriting & translation
- Implementation & performance evaluation
- Conclusions

Motivation

Cancer researcher interested in the changes in chromosome 17 (associated with prostate, bladder, breast cancers) wants to

find single nucleotide polymorphisms (SNPs) associated with chromosome 17

Motivation

Cancer researcher interested in the changes in chromosome 17 (associated with prostate, bladder, breast cancers) wants to

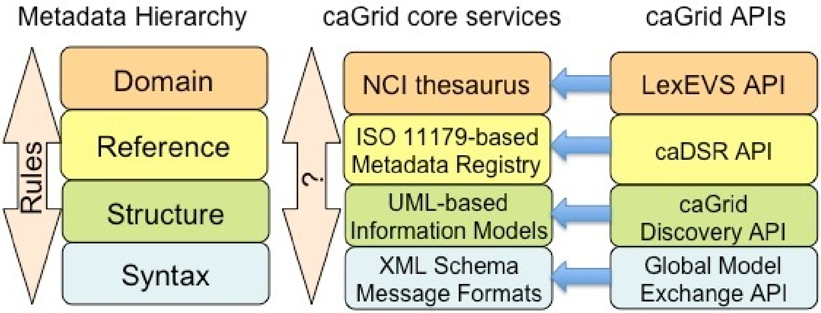
find single nucleotide polymorphisms (SNPs) associated with chromosome 17

Software infrastructures to manage and analyse cancer data from heterogeneous data sources

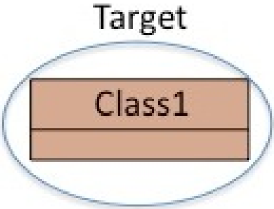
- UK National Cancer Research Institute (NCRI) Informatics Initiative: ONcology Information eXchange (ONIX)
- US National Cancer Institute (NCI) caBIG® programme: caGrid infrastructure

caGrid Infrastructure

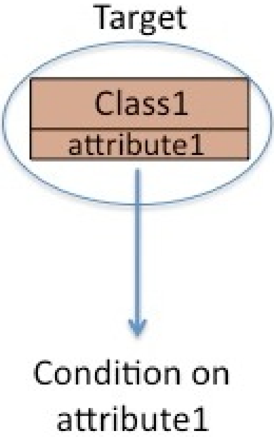
Service-oriented, model-driven infrastructure.



caGrid Query Language (CQL)



caGrid Query Language (CQL)



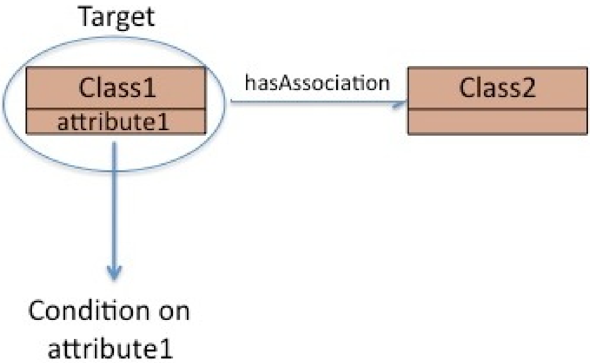
Target

Class1

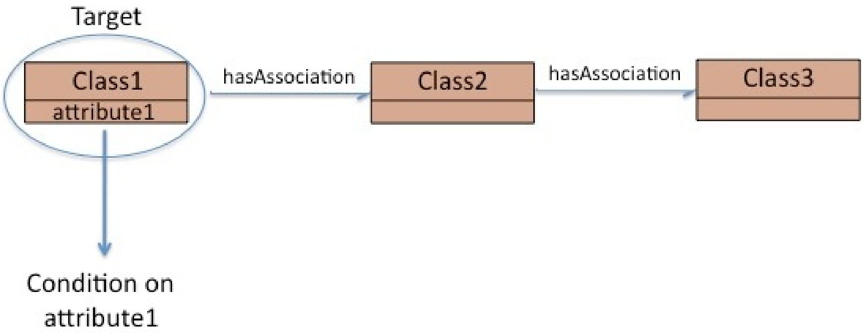
attribute1

Condition on attribute1

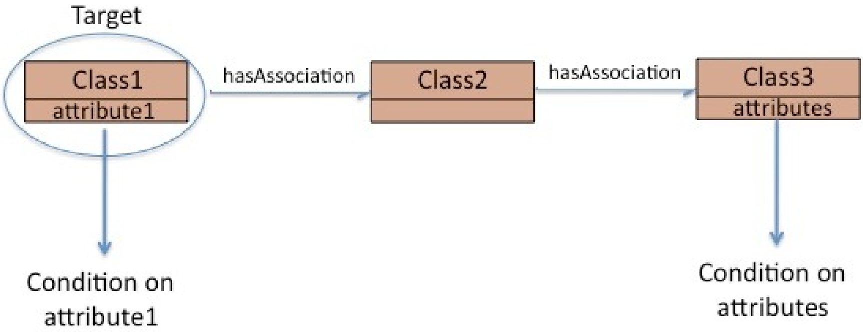
caGrid Query Language (CQL)



caGrid Query Language (CQL)



caGrid Query Language (CQL)



Target

Class1

attribute1

hasAssociation

Class2

hasAssociation

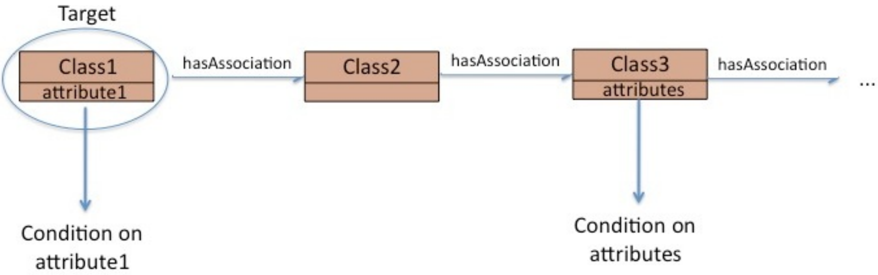
Class3

attributes

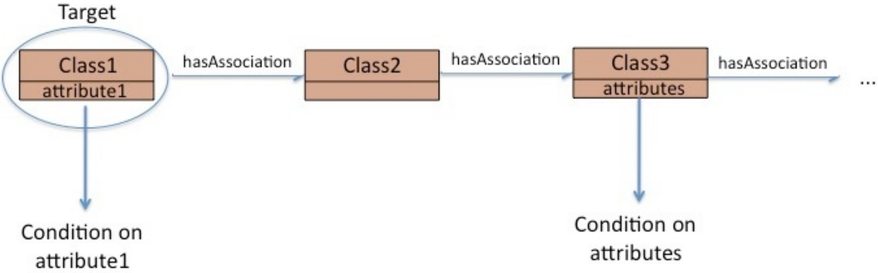
Condition on attribute1

Condition on attributes

caGrid Query Language (CQL)



caGrid Query Language (CQL)



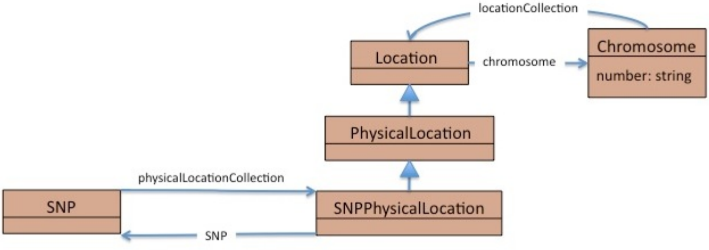
Navigational (path-finding) query language over the structure of caGrid data resources

Motivation — revisited

caBIO data service (cancer Bioinformatics Infrastructure Objects) — biomedical data from a variety of curated data sources

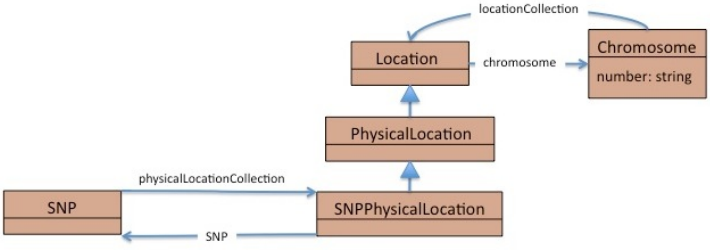
Motivation — revisited

caBIO data service (cancer Bioinformatics Infrastructure Objects) — biomedical data from a variety of curated data sources



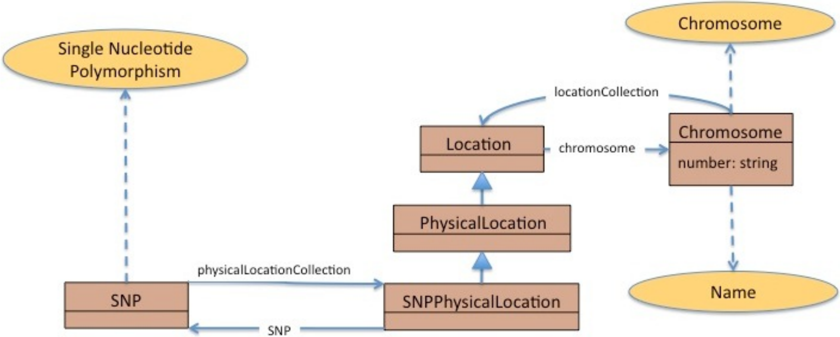
Motivation — revisited

caBIO data service (cancer Bioinformatics Infrastructure Objects) — biomedical data from a variety of curated data sources

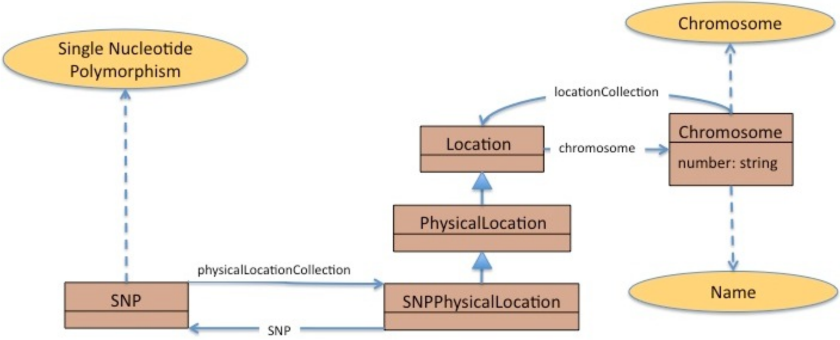


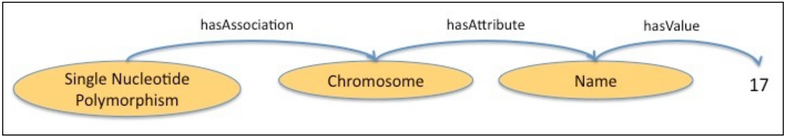
```
<ns1:CQLQuery xmlns:ns1="http://CQL.caBIG/1/gov.nih.nci.cagrid.CQLQuery">
 <ns1:Target name="gov.nih.nci.cabio.domain.SNP">
   <ns1:Association name="gov.nih.nci.cabio.domain.SNPPhysicalLocation"
   roleName="physicalLocationCollection">
    <ns1:Association name="gov.nih.nci.cabio.domain.Chromosome"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

Objective: ontology-based queries

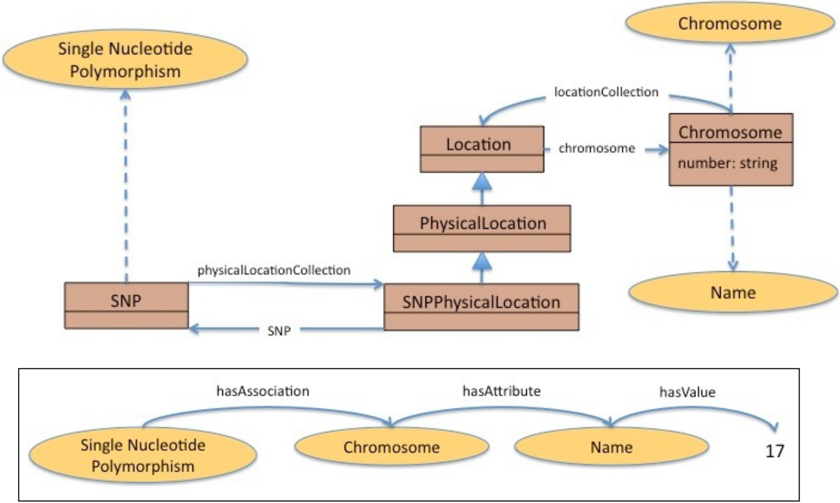


Objective: ontology-based queries





Objective: ontology-based queries

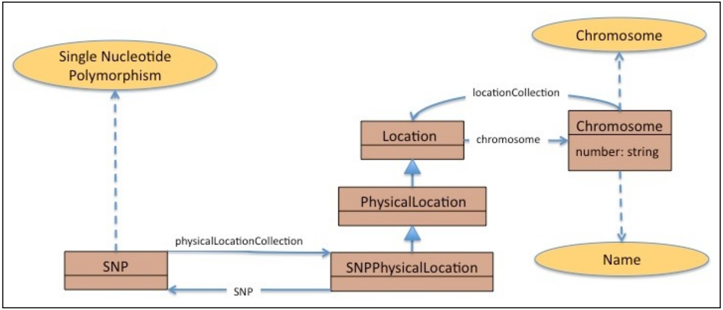


Semantic layer on top of caGrid structural layer

Overview

- Motivation: queries over cancer data
- Background: caGrid infrastructure
  - caGrid query language (CQL)
- Objective: ontology-based queries over the caGrid infrastructure
- **Approach:**
  - **OWL representation of caGrid models**
  - **Query rewriting & translation**
- Implementation & performance evaluation
- Conclusions

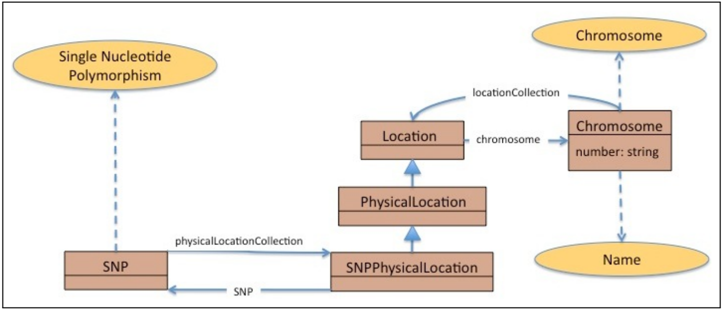
OWL representation of caGrid models



UML class diagrams

$$
\begin{aligned}
\text{c:Chromosome} &\sqsubseteq \text{u:UMLClass}\\
\text{c:number} &\sqsubseteq \text{u:UMLAttribute}\\
\text{c:number} &\sqsubseteq \exists\text{u:hasValue.xsd:string}\\
\text{c:locationCollection} &\sqsubseteq \text{u:hasAssociation}
\end{aligned}
$$

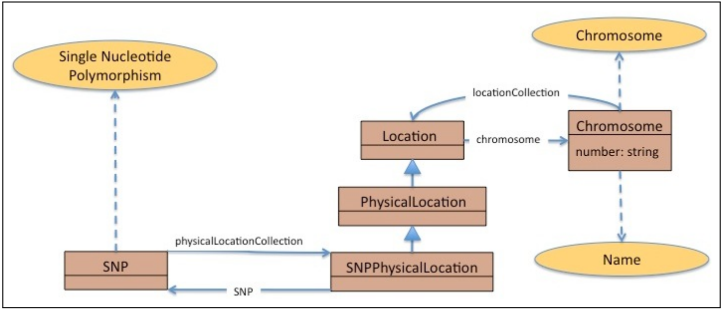
OWL representation of caGrid models



UML class diagrams

$$
\begin{aligned}
\text{c:PhysicalLocation} &\sqsubseteq \text{c:Location}\\
\text{c:Chromosome} &\sqsubseteq \exists\,\text{c:locationCollection.c:Location}\\
\text{c:Chromosome} &\sqsubseteq \exists\,\text{u:hasAttribute.c:number}\\
\text{c:PhysicalLocation} &\sqsubseteq \exists\,\text{c:chromosome.c:Chromosome}
\end{aligned}
$$

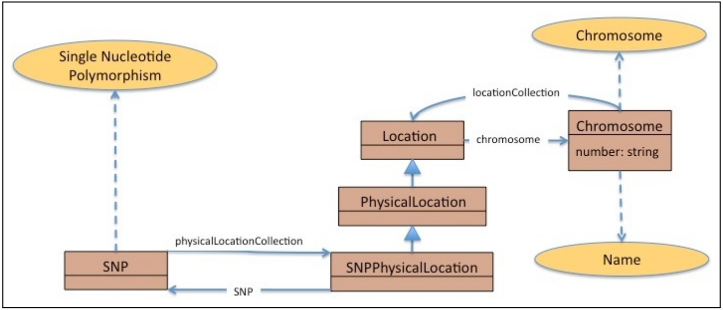
OWL representation of caGrid models



Semantic annotations

$$\begin{aligned} \text{c:SNP} &\sqsubseteq \text{n:Single\_Nucleotide\_Polymorphism} \\ \text{c:Chromosome} &\sqsubseteq \text{n:Chromosome} \\ \text{c:number} &\sqsubseteq \text{n:Name} \end{aligned}$$

OWL representation of caGrid models



Module extraction from NCIt

- Each caGrid information model refers to a subset Σ of the NCIt vocabulary — *relevant* terms and relationships
- NCIt *module* for each data model: Logic-based *module* extraction

Query rewriting and translation

ontology-based query →CQL

Query rewriting and translation

Parsing

n:Single_ Nucleotide_ Polymorphism and hasAssociation some (n:Chromosome and hasAttribute some (n:Name and hasValue value "17"))

Query rewriting and translation

Parsing

n:Single_ Nucleotide_ Polymorphism and hasAssociation some (n:Chromosome and hasAttribute some (n:Name and hasValue value "17"))

UML Extraction

c:SNP and hasAssociation some (c:Chromosome and hasAttribute some (c:number and hasValue value "17"))

Query rewriting and translation

Parsing

n:Single_ Nucleotide_ Polymorphism and hasAssociation some (n:Chromosome and hasAttribute some (n:Name and hasValue value "17"))

UML Extraction

c:SNP and hasAssociation some (c:Chromosome and hasAttribute some (c:number and hasValue value "17"))

Data Values Extraction

c:SNP and hasAssociation some (c:Chromosome and hasAttribute some (c:number))

Query rewriting and translation

Parsing

n:Single_ Nucleotide_ Polymorphism and hasAssociation some (n:Chromosome and hasAttribute some (n:Name and hasValue value "17"))

UML Extraction

c:SNP and hasAssociation some (c:Chromosome and hasAttribute some (c:number and hasValue value "17"))

Data Values Extraction

c:SNP and hasAssociation some (c:Chromosome and hasAttribute some (c:number))

Semantic Validation

Query satisfiable in the ontology?

Query rewriting and translation

Properties Path Finder

c:SNP and c:physicalLocationCollection some c:SNPPhysicalLocation and c:chromosome some (c:Chromosome and hasAttribute some (c:number))

Query rewriting and translation

Properties Path Finder

c:SNP and c:physicalLocationCollection some c:SNPPhysicalLocation and c:chromosome some (c:Chromosome and hasAttribute some (c:number))

Data Values Addition

c:SNP and c:physicalLocationCollection some c:SNPPhysicalLocation and c:chromosome some (c:Chromosome and hasAttribute some (c:number and hasValue value "17"))

Query rewriting and translation

Properties Path Finder

c:SNP and c:physicalLocationCollection some c:SNPPhysicalLocation and c:chromosome some (c:Chromosome and hasAttribute some (c:number))

Data Values Addition

c:SNP and c:physicalLocationCollection some c:SNPPhysicalLocation and c:chromosome some (c:Chromosome and hasAttribute some (c:number and hasValue value "17"))

OWL Expression to MCC Translation

$\uplus \{ s \, ▯ \, s \leftarrow SNP, \; r \leftarrow s.physicalLocationCollection, \; r \leftarrow SNPPhysicalLocation, \; c \leftarrow r.chromosome, \; c \leftarrow Chromosome, \; c.number{=}17 \}$

Query rewriting and translation

Properties Path Finder

c:SNP and c:physicalLocationCollection some c:SNPPhysicalLocation and c:chromosome some (c:Chromosome and hasAttribute some (c:number))

Data Values Addition

c:SNP and c:physicalLocationCollection some c:SNPPhysicalLocation and c:chromosome some (c:Chromosome and hasAttribute some (c:number and hasValue value "17"))

OWL Expression to MCC Translation

⊎ *{ s ▯ s ← SNP, r ← s.physicalLocationCollection, r ← SNPPhysicalLocation, c ← r.chromosome, c ← Chromosome, c.number=17 }*

MCC to CQL Translation

CQL Query

Implementation & performance evaluation

- Two modules: OWL generator (exposed as a caGrid analytical service) & query rewriting/translation
- Java, caGrid 1.3, OWLAPI 3.1, Pellet 2.2.2, HermiT 1.3.0

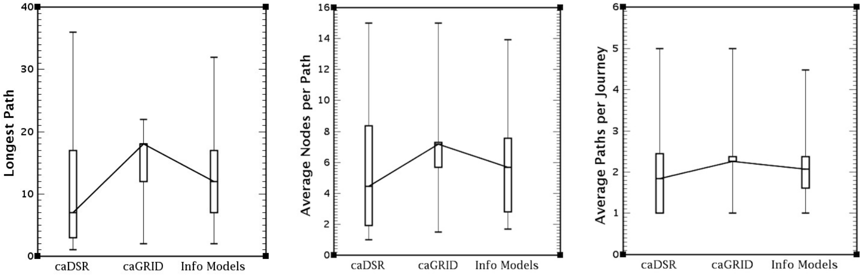
Implementation & performance evaluation

- Two modules: OWL generator (exposed as a caGrid analytical service) & query rewriting/translation
- Java, caGrid 1.3, OWLAPI 3.1, Pellet 2.2.2, HermiT 1.3.0
- Performance
  - Analysis of generated OWL ontologies (caGrid models) — path metrics
  - Ontology generation, module extraction & classification
  - Query rewriting/translation

Implementation & performance evaluation

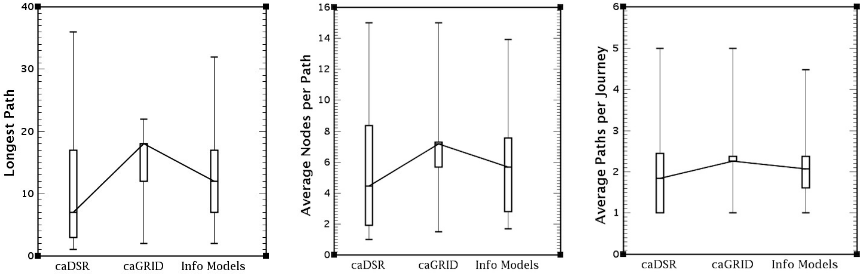
- Two modules: OWL generator (exposed as a caGrid analytical service) & query rewriting/translation
- Java, caGrid 1.3, OWLAPI 3.1, Pellet 2.2.2, HermiT 1.3.0
- Performance
  - Analysis of generated OWL ontologies (caGrid models) — path metrics
  - Ontology generation, module extraction & classification
  - Query rewriting/translation
- Three groups of caGrid models
  - caDSR — registered in caDSR
  - caGrid — registered in caGrid index service
  - InfoModels — models supported by deployed services

Analysis of OWL representation



Path Metrics

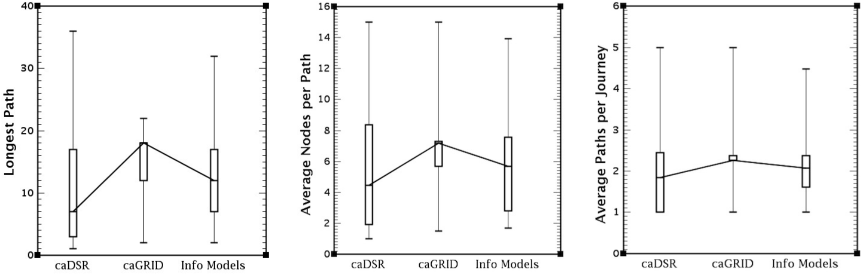
Analysis of OWL representation



Path Metrics

- Longest path: up to 36 nodes; for 75 % of the projects in each category their length is less than 17 or 18

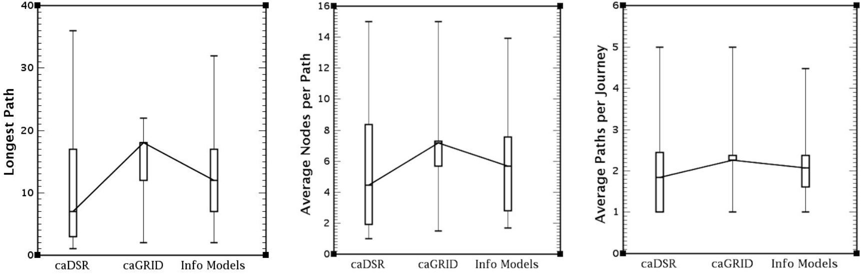
Analysis of OWL representation



Path Metrics

- Longest path: up to 36 nodes; for 75 % of the projects in each category their length is less than 17 or 18
- Average path length: median between 4 and 7 nodes; for 75 % of the InfoModels it is less than 8

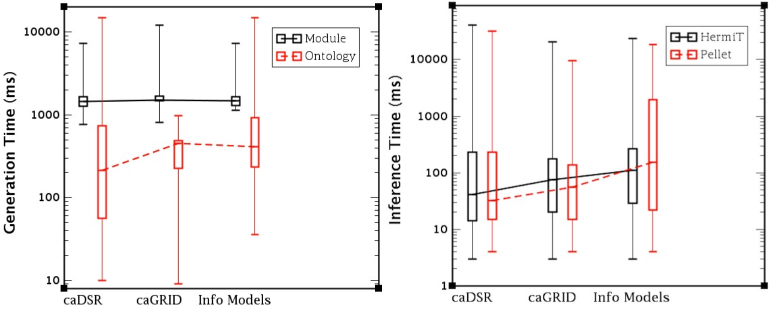
Analysis of OWL representation



Path Metrics

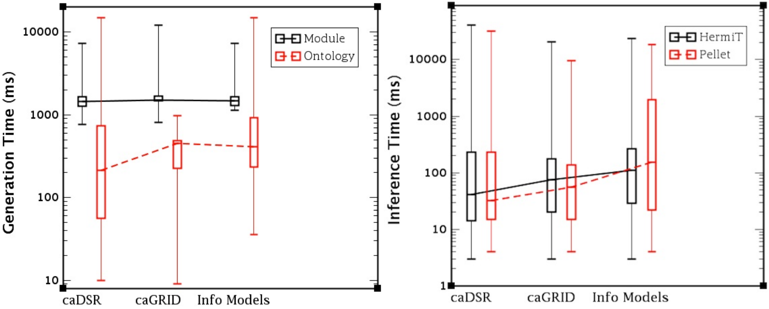
- Longest path: up to 36 nodes; for 75 % of the projects in each category their length is less than 17 or 18
- Average path length: median between 4 and 7 nodes; for 75 % of the InfoModels it is less than 8
- Average paths per journey: median ~ 2 paths per journey; for 75 % of the projects (3 categories), less than 2.5

Ontologies, modules & classification



Generation and inference times

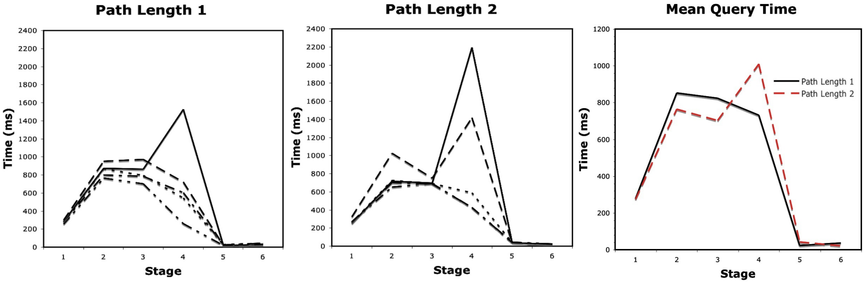
Ontologies, modules & classification



Generation and inference times

- 75 % of NCIt modules, extraction takes less than 2 seconds & even less time for ontology generation
- median inference time (Pellet & HermiT reasoners): less than 100 ms

Query rewriting/translation evaluation



Query rewriting — path lengths 1 and 2, and mean values

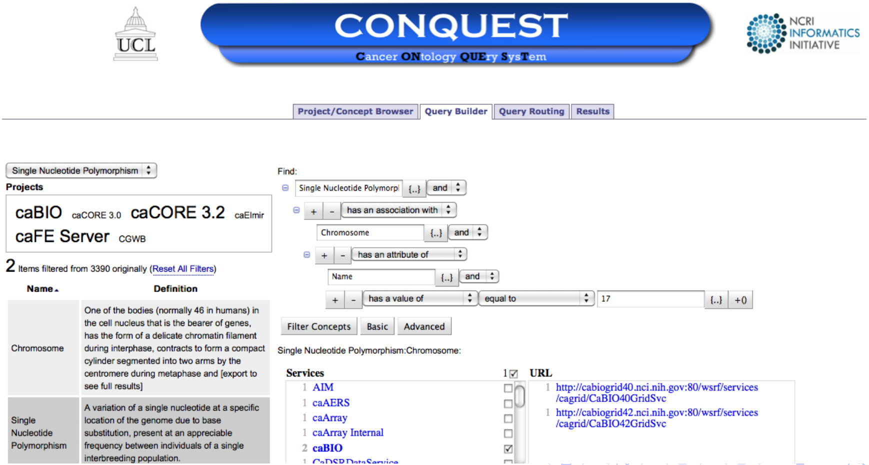
- Stages: (1) parsing, (2) UML extraction, (3) validation, (4) path finding, (5) MCC conversion and (6) CQL conversion
- Path length: affects path-finding stage, rest of stages remain largely unaffected.

Work in progress...

- Explore OWL2EL reasoners — improve path finding stage
- Building a query suite
- GUI development

Work in progress...

- Explore OWL2EL reasoners — improve path finding stage
- Building a query suite
- GUI development



UCL

CONQUEST

Cancer ONtology QUEry SysTem

NCRI INFORMATICS INITIATIVE

Single Nucleotide Polymorphism

Projects

caBIO caCORE 3.0 caCORE 3.2 caElmir caFE Server CGWB

2 Items filtered from 3390 originally (Reset All Filters)

Name	Definition
Chromosome	One of the bodies (normally 46 in humans) in the cell nucleus that is the bearer of genes, has the form of a delicate chromatin filament during interphase, contracts to form a compact cylinder segmented into two arms by the centromere during metaphase and [export to see full results]
Single Nucleotide Polymorphism	A variation of a single nucleotide at a specific location of the genome due to base substitution, present at an appreciable frequency between individuals of a single interbreeding population.

Find:

Single Nucleotide Polymorpl {..} and

+ - has an association with

Chromosome {..} and

+ - has an attribute of

Name {..} and

+ - has a value of equal to 17 {..} +()

Single Nucleotide Polymorphism:Chromosome:

Services 1☑

1 AIM □
1 caAERS □
1 caArray □
1 caArray Internal □
2 **caBIO** ☑
1 CaDSRDataService □

URL

1 http://cabiogrid40.nci.nih.gov:80/wsrf/services/cagrid/CaBIO40GridSvc
1 http://cabiogrid42.nci.nih.gov:80/wsrf/services/cagrid/CaBIO42GridSvc

Conclusions

- Ontology-based queries over caGrid: design & implementation
- General approach for service-oriented, model-driven infrastructure — only last step of query rewriting (MCC2CQL) depends on caGrid
- Generation of OWL2 ontologies from annotated UML models (ISO11179 standard)
- Analysis of generated ontologies — path metric
- caGrid analytical service for the OWL generator
- Analysis of CQL
- Query rewriting/translation procedure — OWL class expressions →MCC →CQL
- Performance evaluation — OWL generation, module extraction, classification
- Assessment of query rewriting/translation procedure and its viability

Conclusions

- Ontology-based queries over caGrid: design & implementation
- General approach for service-oriented, model-driven infrastructure — only last step of query rewriting (MCC2CQL) depends on caGrid
- Generation of OWL2 ontologies from annotated UML models (ISO11179 standard)
- Analysis of generated ontologies — path metric
- caGrid analytical service for the OWL generator
- Analysis of CQL
- Query rewriting/translation procedure — OWL class expressions →MCC →CQL
- Performance evaluation — OWL generation, module extraction, classification
- Assessment of query rewriting/translation procedure and its viability

Conclusions

- Ontology-based queries over caGrid: design & implementation
- General approach for service-oriented, model-driven infrastructure — only last step of query rewriting (MCC2CQL) depends on caGrid
- Generation of OWL2 ontologies from annotated UML models (ISO11179 standard)
- Analysis of generated ontologies — path metric
- caGrid analytical service for the OWL generator
- Analysis of CQL
- Query rewriting/translation procedure — OWL class expressions →MCC →CQL
- Performance evaluation — OWL generation, module extraction, classification
- Assessment of query rewriting/translation procedure and its viability

Conclusions

- Ontology-based queries over caGrid: design & implementation
- General approach for service-oriented, model-driven infrastructure — only last step of query rewriting (MCC2CQL) depends on caGrid
- Generation of OWL2 ontologies from annotated UML models (ISO11179 standard)
- Analysis of generated ontologies — path metric
- caGrid analytical service for the OWL generator
- Analysis of CQL
- Query rewriting/translation procedure — OWL class expressions →MCC →CQL
- Performance evaluation — OWL generation, module extraction, classification
- Assessment of query rewriting/translation procedure and its viability

Conclusions

- Ontology-based queries over caGrid: design & implementation
- General approach for service-oriented, model-driven infrastructure — only last step of query rewriting (MCC2CQL) depends on caGrid
- Generation of OWL2 ontologies from annotated UML models (ISO11179 standard)
- Analysis of generated ontologies — path metric
- caGrid analytical service for the OWL generator
- Analysis of CQL
- Query rewriting/translation procedure — OWL class expressions →MCC →CQL
- Performance evaluation — OWL generation, module extraction, classification
- Assessment of query rewriting/translation procedure and its viability

Conclusions

- Ontology-based queries over caGrid: design & implementation
- General approach for service-oriented, model-driven infrastructure — only last step of query rewriting (MCC2CQL) depends on caGrid
- Generation of OWL2 ontologies from annotated UML models (ISO11179 standard)
- Analysis of generated ontologies — path metric
- caGrid analytical service for the OWL generator
- Analysis of CQL
- Query rewriting/translation procedure — OWL class expressions →MCC →CQL
- Performance evaluation — OWL generation, module extraction, classification
- Assessment of query rewriting/translation procedure and its viability

Conclusions

- Ontology-based queries over caGrid: design & implementation
- General approach for service-oriented, model-driven infrastructure — only last step of query rewriting (MCC2CQL) depends on caGrid
- Generation of OWL2 ontologies from annotated UML models (ISO11179 standard)
- Analysis of generated ontologies — path metric
- caGrid analytical service for the OWL generator
- Analysis of CQL
- Query rewriting/translation procedure — OWL class expressions →MCC →CQL
- Performance evaluation — OWL generation, module extraction, classification
- Assessment of query rewriting/translation procedure and its viability

Conclusions

- Ontology-based queries over caGrid: design & implementation
- General approach for service-oriented, model-driven infrastructure — only last step of query rewriting (MCC2CQL) depends on caGrid
- Generation of OWL2 ontologies from annotated UML models (ISO11179 standard)
- Analysis of generated ontologies — path metric
- caGrid analytical service for the OWL generator
- Analysis of CQL
- Query rewriting/translation procedure — OWL class expressions →MCC →CQL
- Performance evaluation — OWL generation, module extraction, classification
- Assessment of query rewriting/translation procedure and its viability

Conclusions

- Ontology-based queries over caGrid: design & implementation
- General approach for service-oriented, model-driven infrastructure — only last step of query rewriting (MCC2CQL) depends on caGrid
- Generation of OWL2 ontologies from annotated UML models (ISO11179 standard)
- Analysis of generated ontologies — path metric
- caGrid analytical service for the OWL generator
- Analysis of CQL
- Query rewriting/translation procedure — OWL class expressions →MCC →CQL
- Performance evaluation — OWL generation, module extraction, classification
- Assessment of query rewriting/translation procedure and its viability

Acknowledgments

- National Cancer Research Institute (NCRI) Informatics Initiative, UK
- Konrad Rokicki, SAIC/NCI CBIIT

Thank you!

Thank you!

Questions?