Towards ontology-driven querying and integration of cancer-research data

Alejandra González-Beltrán
University College London

Joshua Phillips
SemanticBits
What do we want to do?
Case study

A scientist interested in the gene breast cancer 1, early onset (BRCA1) wants to retrieve information about:

- the single nucleotide polymorphisms (SNPs) associated with the gene
- the pathways the gene is involved in, where the gene encodes proteins with ligand binding domains that are located within a certain sequence range
What do we want to do?

Case study

A scientist interested in the gene *breast cancer 1, early onset* (*BRCA1*) wants to retrieve information about:

- the single nucleotide polymorphisms (SNPs) associated with the gene
- the pathways the gene is involved in, where the gene encodes proteins with ligand binding domains that are located within a certain sequence range

Where to find this information?

Have we found what we meant?

Can we search in terms of concepts and relations?

Can we work with the results coming from various sources?
Context

- **Where to find this information?**
  - caBIG® programme & NCRI Informatics Initiative
  - caGrid portal & ONIX
  - Platforms for data sharing, where data is made available through open, standards-based Web Services

- **Have we found what we meant?**
  - Data sets precisely described – unambiguous meaning through semantic annotations
  - Maximise interoperability of data sets and tools that use them
Objectives

Can we search in terms of concepts and relations?
Can we work with the results coming from various sources?

- Collaboration between caBIG®/NCRI Informatics
- Ontology-based Query WG (caBIG®) / PART2 project (UCL/NCRII/ICL)
  - Support high-level and declarative queries over metadata and data against a unified view of heterogeneous data sources.
  - Support (semantic) data integration through the design and development approaches that are general and applicable to other infrastructures/data
Outline

• Context and objectives
  – Case Study

• Background
  – caGrid overview
  – Semantic web overview

• Ontology-driven approach
  – Concept-based queries
  – Data integration

• Future work

• Conclusions
caGrid overview

caGrid Data Service

Cancer Research Data Source
caGrid overview
caGrid overview

caGrid Data Service

Cancer Research Data Source

Metadata Registry (based on ISO 11179)

EVS

Enterprise Vocabulary Services

Global Model Exchange (GME)

XML-Schema Repository

Annotated UML model

Information Model

RelativeLocation

MarkerRelativeLocation

GeneRelativeLocation

Gene

Gene_Symbol

Single_Nucleotide_Polymorphism

Chromosome

Relative_Value

Location
caGrid overview

caGrid Data Service

CQL query

CQLQuery Processor

Information Model

Cancer Research Data Source

Global Model Exchange (GME) XML-Schema Repository

Metadata Registry (based on ISO 11179)

Annotated UML model

Chromosome

Relative_Value

Location

RelativeLocation

MarkerRelativeLocation

GeneRelativeLocation

Single_Nucleotide_Polymorphism

Gene

Gene_Symbol

Marker

Gene_Symbol/String

Global Model Exchange (GME)
A scientist interested in the gene breast cancer 1, early onset (BRCA1) wants to retrieve the single nucleotide polymorphisms (SNPs) associated with the gene
caGrid overview - caGrid Query Language (CQL)

```xml
<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.GeneRelativeLocation" roleName="relativeLocationCollection">
      <ns1:Association name="gov.nih.nci.cabio.domain.Gene" roleName="gene">
        <ns1:Attribute name="symbol" predicate="EQUAL_TO" value="BRCA1"/>
      </ns1:Association>
    </ns1:Association>
  </ns1:Target>
</ns1:CQLQuery>
```
caGrid overview

- Domain
- Reference
- Structure
- Syntax

Metadata Hierarchy - Pollock'04
caGrid overview

- NCI thesaurus provides a common view
- Making the necessary inferences requires navigating multiple APIs and metadata representations
- **Proposal:** Use a single representation for metadata and data using Semantic web technologies
NCI thesaurus provides a common view

Making the necessary inferences requires navigating multiple APIs and metadata representations

**Proposal:** Use a single representation for metadata and data using Semantic web technologies
caGrid overview

- NCI thesaurus provides a common view
- Making the necessary inferences requires navigating multiple APIs and metadata representations

**Proposal:** Use a single representation for metadata and data using Semantic web technologies
Outline

• Context and objectives
  – Case Study

• Background
  – caGrid overview
  – Semantic web overview

• Ontology-driven approach
  – Concept-based queries
  – Data integration

• Future work

• Conclusions
Semantic web overview

• “... an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation. It is the idea of having data on the Web and linked in a way that it can be used for more effective discovery, automation, integration, and reuse across various applications”

  World Wide Web Consortium (W3C)

• Semantic Grid
  – It captures the relationship between the Grid computing and Semantic Web visions

  Goble et al, IBFI ‘05
Semantic web overview

• “... an extension of the current Web in which information is given **well-defined meaning**, better enabling computers and people to work in cooperation. It is the idea of having **data on the Web** and **linked** in a way that it can be **used** for more effective **discovery, automation, integration, and reuse across various applications**”

  World Wide Web Consortium (W3C)

• **Semantic Grid**
  - It captures the relationship between the Grid computing and Semantic Web visions

  Goble et al, IBFI ‘05
Semantic web overview

- Resource Description Framework (RDF): graph model for describing relations between “things” (resources)
• Resource Description Framework (RDF): graph model for describing relations between “things” (resources)
Semantic web overview

- Resource Description Framework (RDF): graph model for describing relations between “things” (resources)
Semantic web overview

- Resource Description Framework (RDF): graph model for describing relations between “things” (resources)
• Resource Description Framework (RDF): graph model for describing relations between “things” (resources)
• Resource Description Framework (RDF): graph model for describing relations between “things” (resources)
• RDF Schema (RDFS): vocabulary definition language
- Resource Description Framework (RDF): graph model for describing relations between “things” (resources)
- RDF Schema (RDFS): vocabulary definition language

Semantic web overview

cb:Gene
rdfs:type
Resource Description Framework (RDF): graph model for describing relations between “things” (resources)
RDF Schema (RDFS): vocabulary definition language

cb:gene_bcra1
sq:hasAssociation
rdfs:subPropertyOf
cb:gene

cb:SNP
rdfs:type
cb:snp_a

cb:GeneRelativeLocation
rdfs:type
cb:grl_a

cb:Gene
rdfs:type
cb:gene

cb:relativeLocation
sq:hasAssociation
cb:gene

sq:hasValue
“BCRA1”
Semantic web overview

- Resource Description Framework (RDF): graph model for describing relations between “things” (resources)
- RDF Schema (RDFS): vocabulary definition language
- Web Ontology Language (OWL): a more expressive vocabulary definition language – based on Description Logics
Semantic web overview

- Resource Description Framework (RDF): graph model for describing relations between “things” (resources)
- RDF Schema (RDFS): vocabulary definition language
- Web Ontology Language (OWL): a more expressive vocabulary definition language – based on Description Logics
- Ontology: a formal conceptualization of a domain of interest
Semantic web overview

- Resource Description Framework (RDF): graph model for describing relations between “things” (resources)
- RDF Schema (RDFS): vocabulary definition language
- Web Ontology Language (OWL): a more expressive vocabulary definition language – based on Description Logics
- Ontology: a formal conceptualization of a domain of interest
- Formal semantics & inferencing capabilities (reasoning)
Semantic web overview

- Resource Description Framework (RDF): graph model for describing relations between “things” (resources)
- RDF Schema (RDFS): vocabulary definition language
- Web Ontology Language (OWL): a more expressive vocabulary definition language – based on Description Logics
- Ontology: a formal conceptualization of a domain of interest
- Formal semantics & inferencing capabilities (reasoning)
- DL queries and SPARQL queries
Outline

• Context and objectives
  – Case Study

• Background
  – caGrid overview
  – Semantic web/grid overview

• Ontology-driven approach
  – Concept-based queries
  – Data integration

• Future work

• Conclusions
Ontology-driven approach
Ontology-driven approach
Ontology-driven approach
Ontology-driven approach
Annotated UML-to-OWL Generator

Generate an ontology from the data service metadata (annotated UML to OWL transformation)
Query Formulation (User Portal)

Express the concept-based query over the generated ontology

Find objects that have concept *Single_Nucleotide_Polymorphism* and have an association with objects whose concept is *Gene*, which in turn have an attribute with concept *Gene_Symbol*, whose value is “BRCA1”
Query Formulation (User Portal)

Express the concept-based query over the generated ontology

Find objects that have concept *Single_Nucleotide_Polymorphism* and have an association with objects whose concept is *Gene*, which in turn have an attribute with concept *Gene_Symbol*, *whose value is “BRCA1”*

hasConcept some *Single_Nucleotide_Polymorphism* and hasAssociation some (hasConcept some *Gene* and hasAttribute some (hasConcept some *Gene_Symbol* and hasValue value “BRCA1”))

Description Logic query (DL-query) in Manchester OWL Syntax
Query Formulation (User Portal)

Express the concept-based query over the generated ontology

Find objects that have concept *Single_Nucleotide_Polymorphism* and have an association with objects whose concept is *Gene*, which in turn have an attribute with concept *Gene_Symbol*, whose value is “BRCA1”

```
hasConcept some Single_Nucleotide_Polymorphism and hasAssociation some (hasConcept some Gene and hasAttribute some (hasConcept some Gene_Symbol and hasValue value “BRCA1”))
```

Description Logic query (DL-query) in Manchester OWL Syntax

- **Concept-based** query: it can be used for any resource annotated with the same vocabulary
- **High-level** query: it is not based on the structure of a particular target resource
- **Descriptive**: it gives the criteria for the desired data
Semantic Federated Query Processor
Query Transformation

Transform the query using the generated ontology into the caGrid query language

Concept-based query

Syntactic and semantic analysis

Reformulate query in terms of the data service’s elements

Translate DL-query into the Object-query optimisation language (MCC) and normalise

Translate the object-query expression (MCC) to CQL

CQL query

Generated Data Service Ontology
Outline

• Context and objectives
  – Case Study
• Background
  – caGrid overview
  – Semantic web/grid overview
• Ontology-driven approach
  – Concept-based queries
  – Data integration
• Future work
• Conclusions
Ontology-driven approach
Ontology-driven approach
Ontology-driven approach
Back to the case study (2\textsuperscript{nd} part)

- Find pathways that a gene is involved in, where the gene encodes proteins with ligand binding domains that are located within a certain sequence range
Back to the case study (2nd part)

- Find pathways that a gene is involved in, where the gene encodes proteins with ligand binding domains that are located within a certain sequence range

- To find this information, we need to combine data from two data sources (caBIO & gridPIR)
Data integration

“Data integration is the problem of combining data at different sources, and providing the user with a unified view of these data”

Lenzerini PODS ’02

Some current challenges:

• Expressivity of D/CQL is limited – use of solely federated query does not allow sophisticated queries
• Lack of common identifier approach – determining equivalence cannot be automated
metadata and data from caBIO in OWL
metadata and data from gridPIR in OWL
• Rule encoded as SPARQL CONSTRUCT
• Independent of information models
OWL Representation Support
Transformer Service & XML-OWL Transformer
Knowledge-based services
Triple Store & Reasoning
Outline

• Context and objectives
  – Case Study

• Background
  – caGrid overview
  – Semantic web/grid overview

• Ontology-driven approach
  – Concept-based queries
  – Data integration

• Future work

• Conclusions
Future Work

• Extended OWL representation
  – Including caDSR semantics
• Query transformations for multiple resources
• Reasoner Service Interface
  • Align with OLWLink Protocol – standard KB interactions and metadata
• TripleStore Service Interface
  • Align with caBIG® identifiers and Linked Open Data community.
• Support transaction semantics and security
• SPARQL & Linked Data Plug-in for caBIG® Data Services
Conclusions

• caBIG® and NCRI Informatics programmes
• caGrid metadata infrastructure
  – Structural queries
  – Data integration responsibility of the client
• Combination caGrid with Semantic Web technologies
  – Semantic caGrid
• Ontology-driven approach
  • Queries expressed with concepts and relationships
    DL query (reasoning over metadata)
    SPARQL query (reasoning over data & metadata)
• Semantic data integration
• Demonstrated in the caGrid infrastructure
• Generalisable
Further information


Further Information

• Whitepaper
  http://tinyurl.com/qqybmq

• Code (data integration services)
  http://tinyurl.com/cv67of

• Guide to code examples (data integration services)
  http://tinyurl.com/cwv4fy

• Demo for concept-based queries
  http://tinyurl.com/o6uw7z
Acknowledgements

- NCRI Informatics Initiative
- caBIG® programme
- Cancer Research UK
- Prof Anthony Finkelstein (University College London)
- Dr J. Max Wilkinson (British Library)
- Prof Jeff Kramer (Imperial College London)
- Jim McCusker (Yale University)
- Michael Krauthammer (Yale University)
- Denise Warzel (CBIIT, National Cancer Institute)
Thank you!