Domain Concept-Based Queries for Cancer Research Data Sources

Alejandra González Beltrán

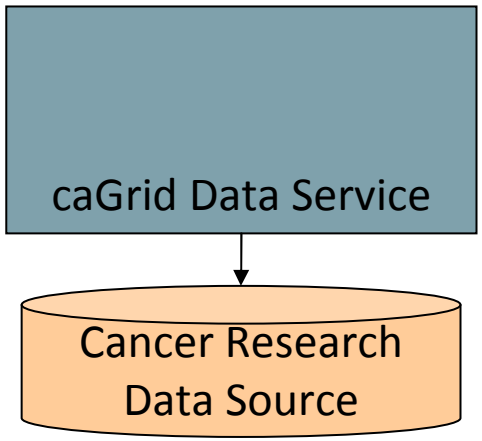
Joint work with Anthony Finkelstein (UCL), J Max Wilkinson (NCRI) and Jeff Kramer (ICL)

CBMS 2009
The 22nd IEEE International Symposium on Computer-Based Medical Systems
August 3-4 2009, Albuquerque, New Mexico, USA

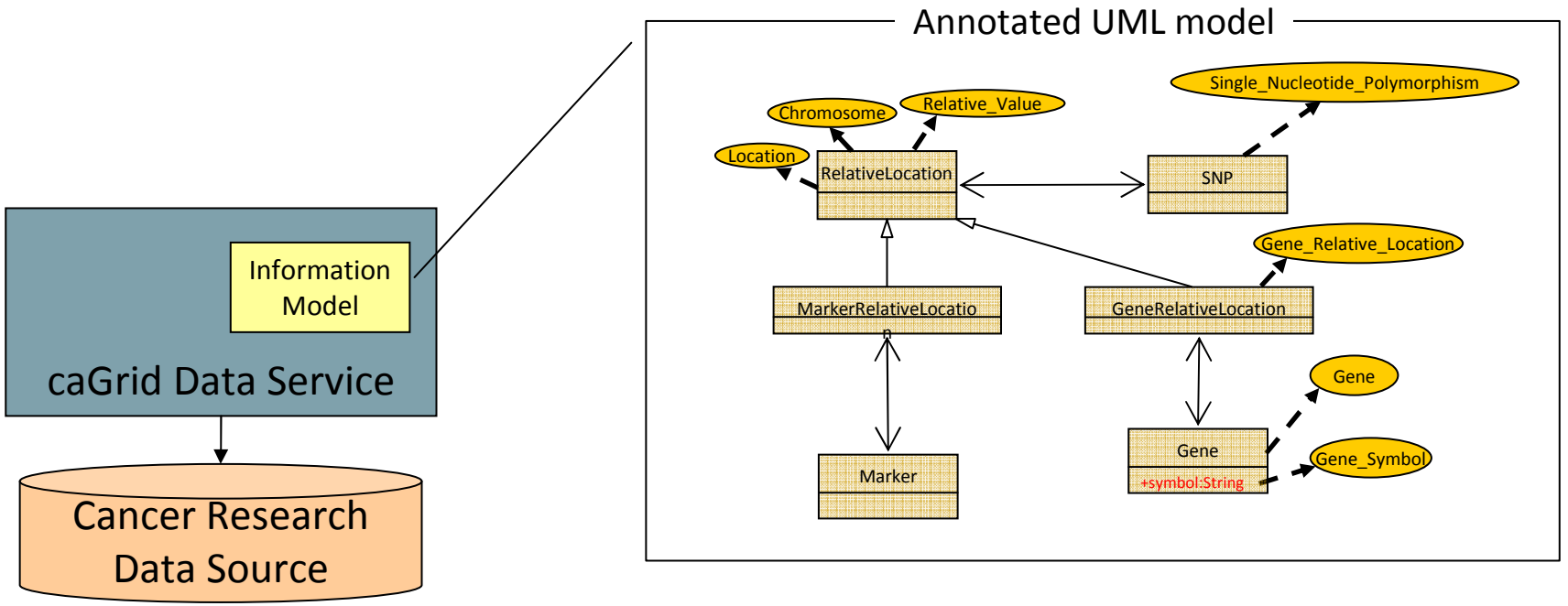
Overview

- ONIX (ONcology Information eXchange) – UK NCRI platform to facilitate access to distributed cancer-research data sources
  - Interoperability with the caGrid infrastructure
  - Support for non-caGrid resources
- caGrid metadata infrastructure and query language
  - Queries based on the structure of the resource
  - No support for concept-based queries
- Goal: support for high-level and descriptive queries of cancer-research data sources, expressed using domain concepts and their relationships
- Architecture & Approach: caGrid + Semantic Web technologies
  - Concept-based queries and data integration
  - Extensible to non-caGrid resources
- Conclusions

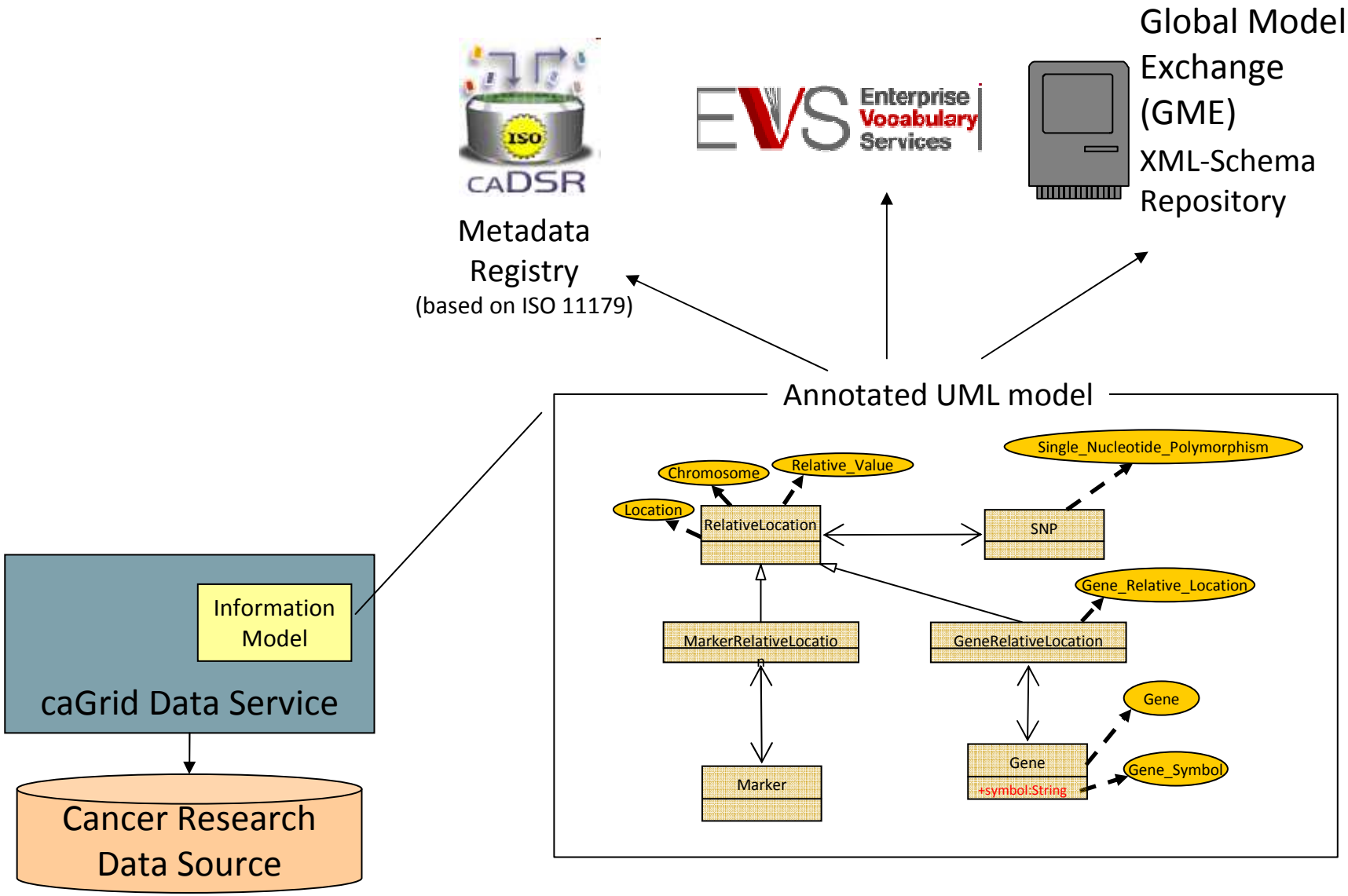
caGrid infrastructure



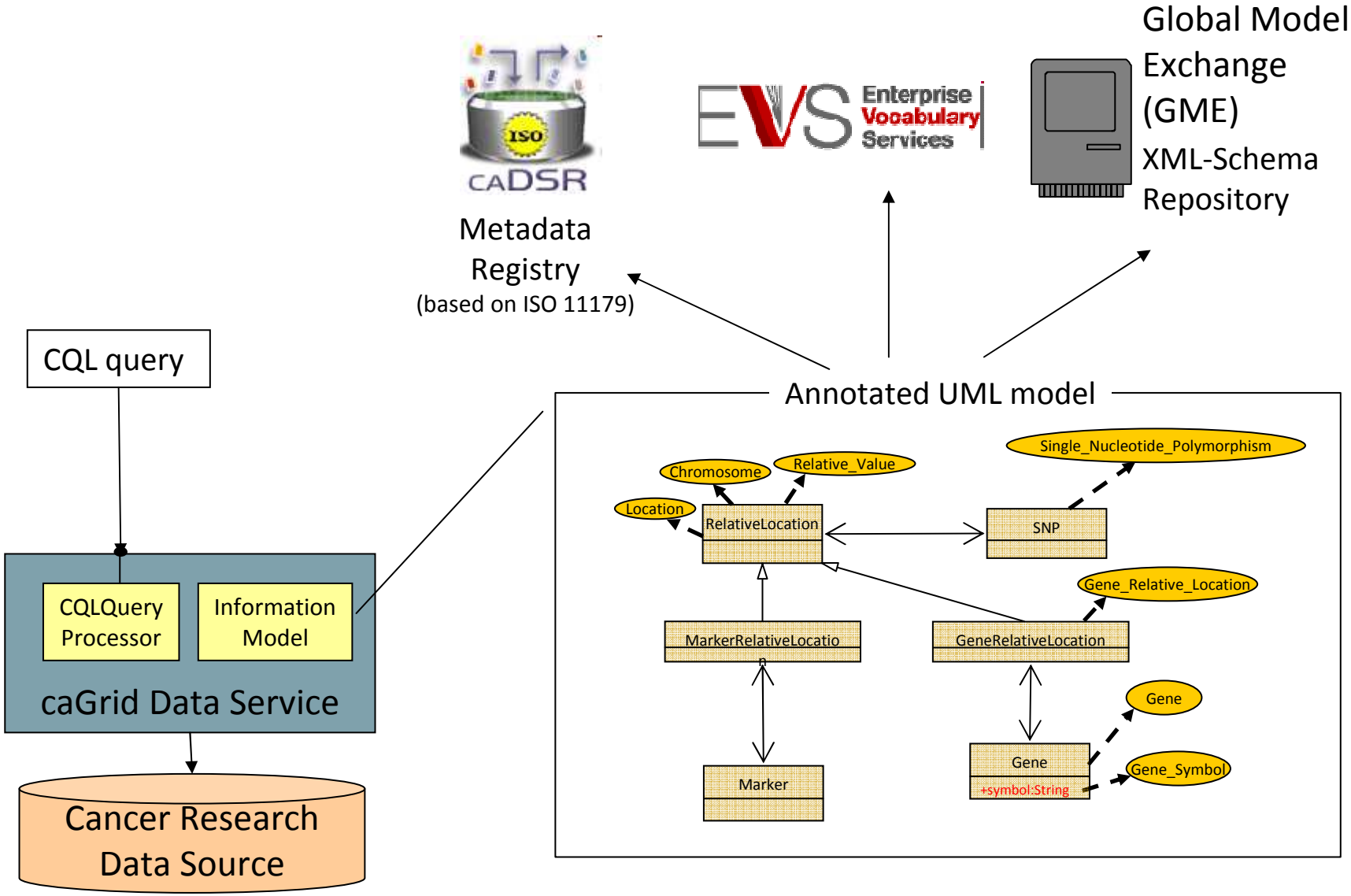
caGrid infrastructure



caGrid infrastructure



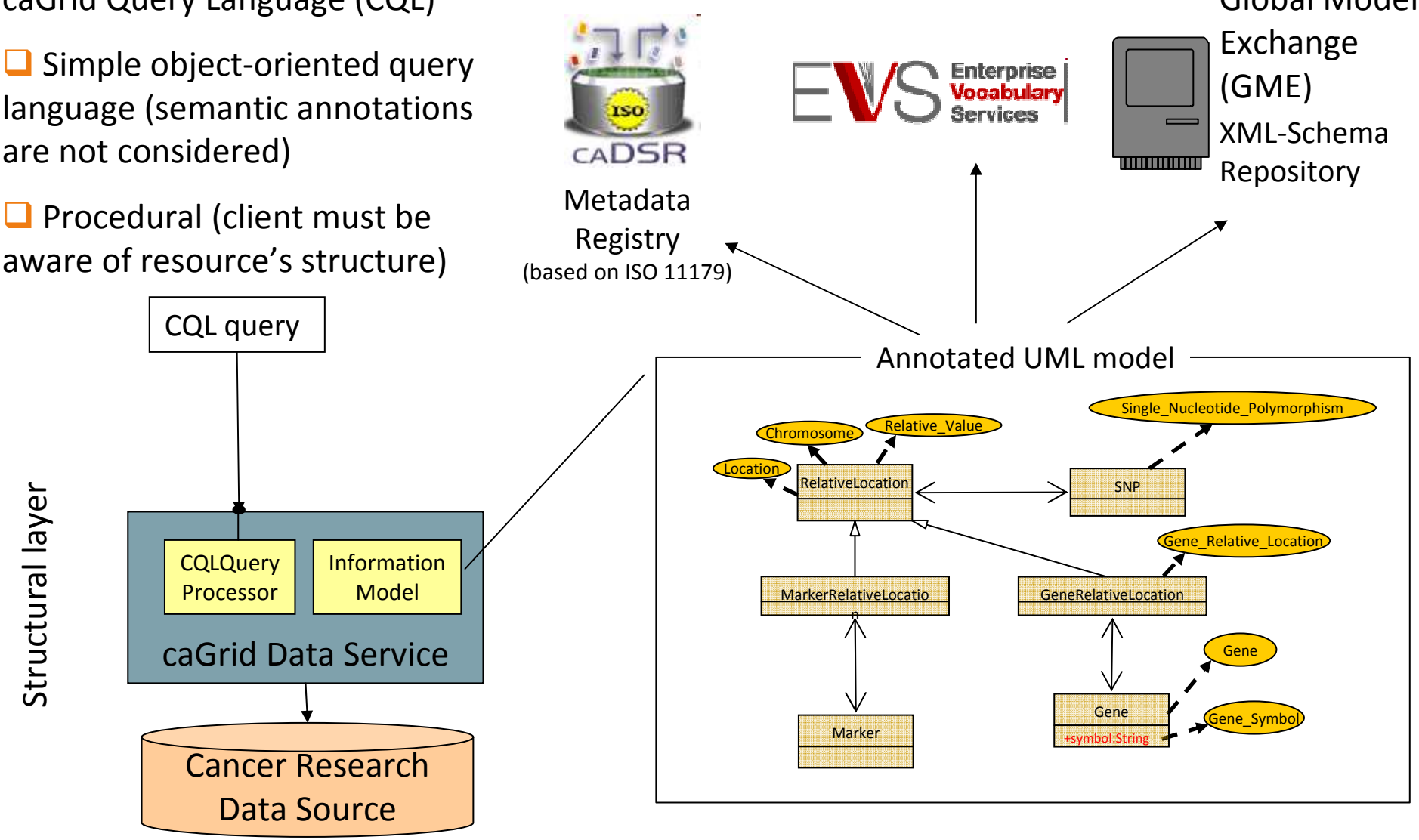
caGrid infrastructure



caGrid infrastructure

caGrid Query Language (CQL)

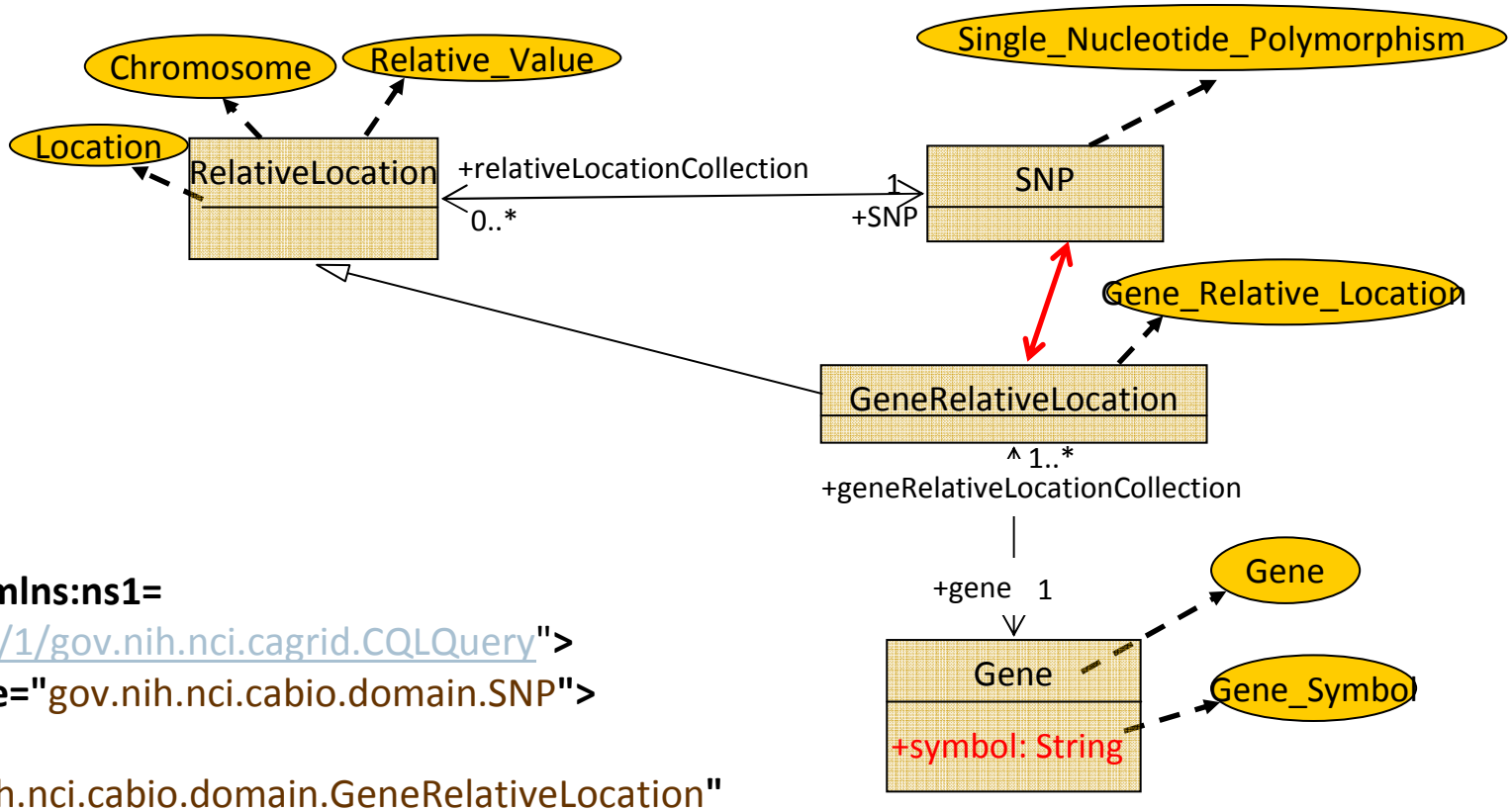
- Simple object-oriented query language (semantic annotations are not considered)
- Procedural (client must be aware of resource's structure)



Motivating example

- A biomedical scientist is interested in finding single nucleotide polymorphisms (SNPs) associated with the gene *Transforming Growth Factor Beta 1* (*TGFB1*)

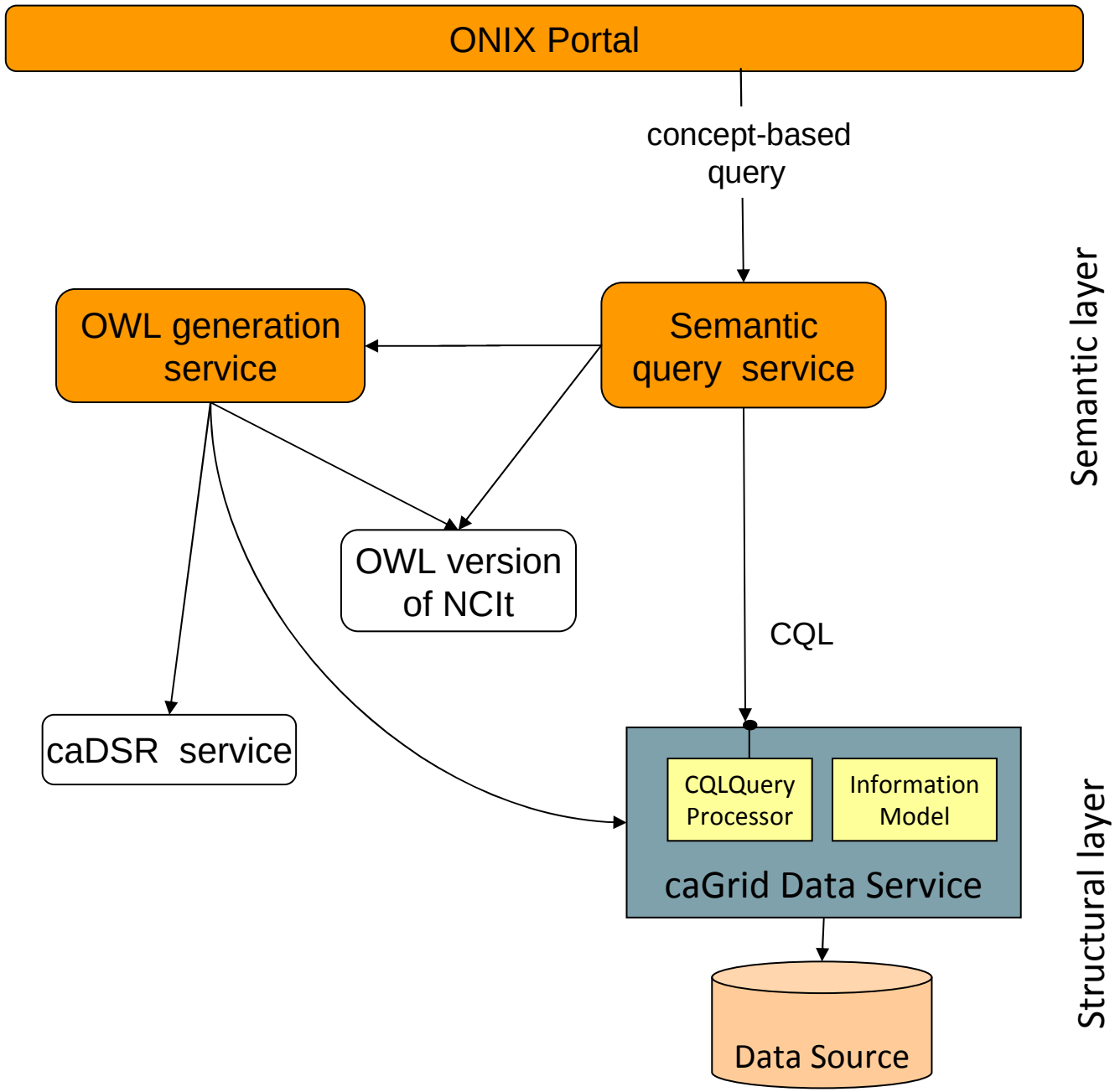
caGrid query language (CQL)




```
<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.GeneRelativeLocation"
      roleName="relativeLocationCollection">
   <ns1:Association  name="gov.nih.nci.cabio.domain.Gene“
      roleName="gene">
    <ns1:Attribute name="symbol" predicate= "EQUAL_TO" value="TGFB1"/>
   </ns1:Association>  </ns1:Association> </ns1:Target></ns1:CQLQuery>
```

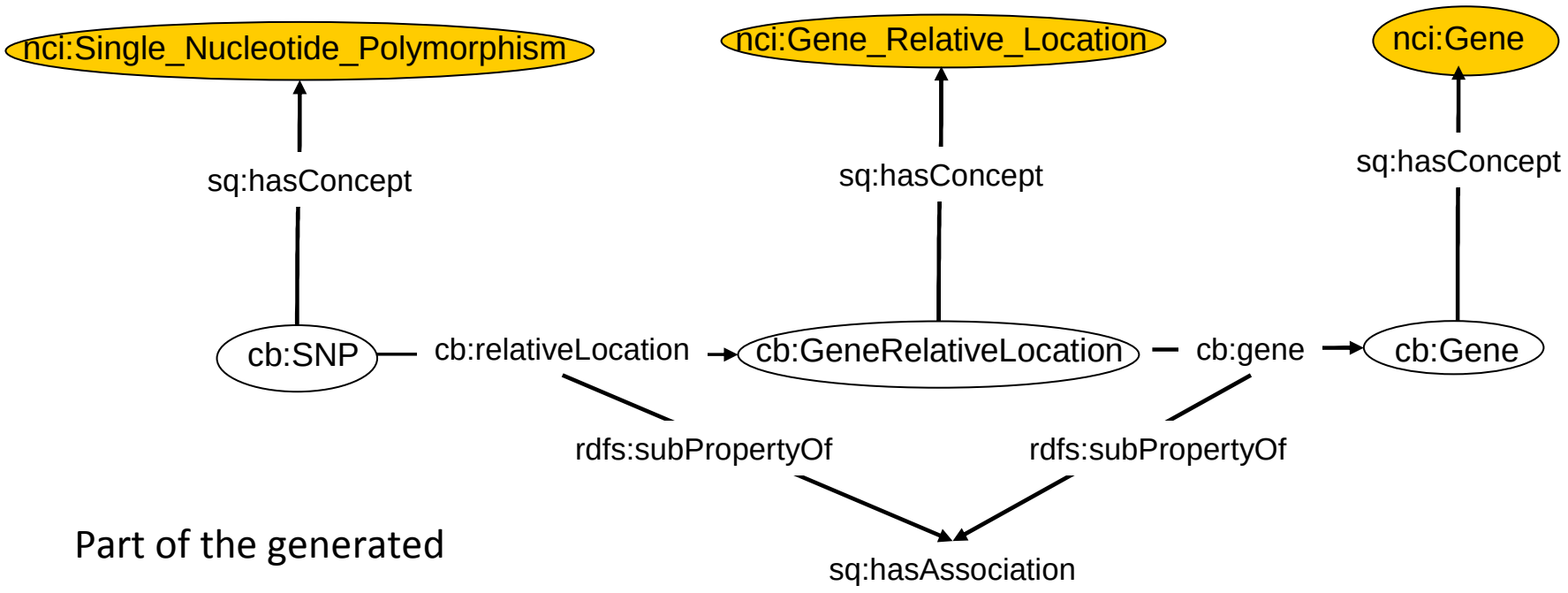
Architecture

- Web Ontology Language (OWL): W3C recommendation for knowledge representation
- Reasoning: inference capabilities
- OWL generation service: develops OWL ontologies from information models
- Semantic query service: transforms concept-based queries to CQL using the generated ontologies + reasoning
- Prototype: OWLAPI and Pellet



Approach

1) Generate an ontology from the data service metadata (annotated UML to OWL transformation)



Part of the generated ontology

Approach

2) 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 "TGFB1"*

Approach

2) 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 "TGFB1"*

- hasConcept some *Single_Nucleotide_Polymorphism* and hasAssociation some (hasConcept some *Gene* and hasAttribute some (hasConcept some *Gene_Symbol* and hasValue value "*TGFB1*")

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

Approach

2) 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 "TGFB1"*

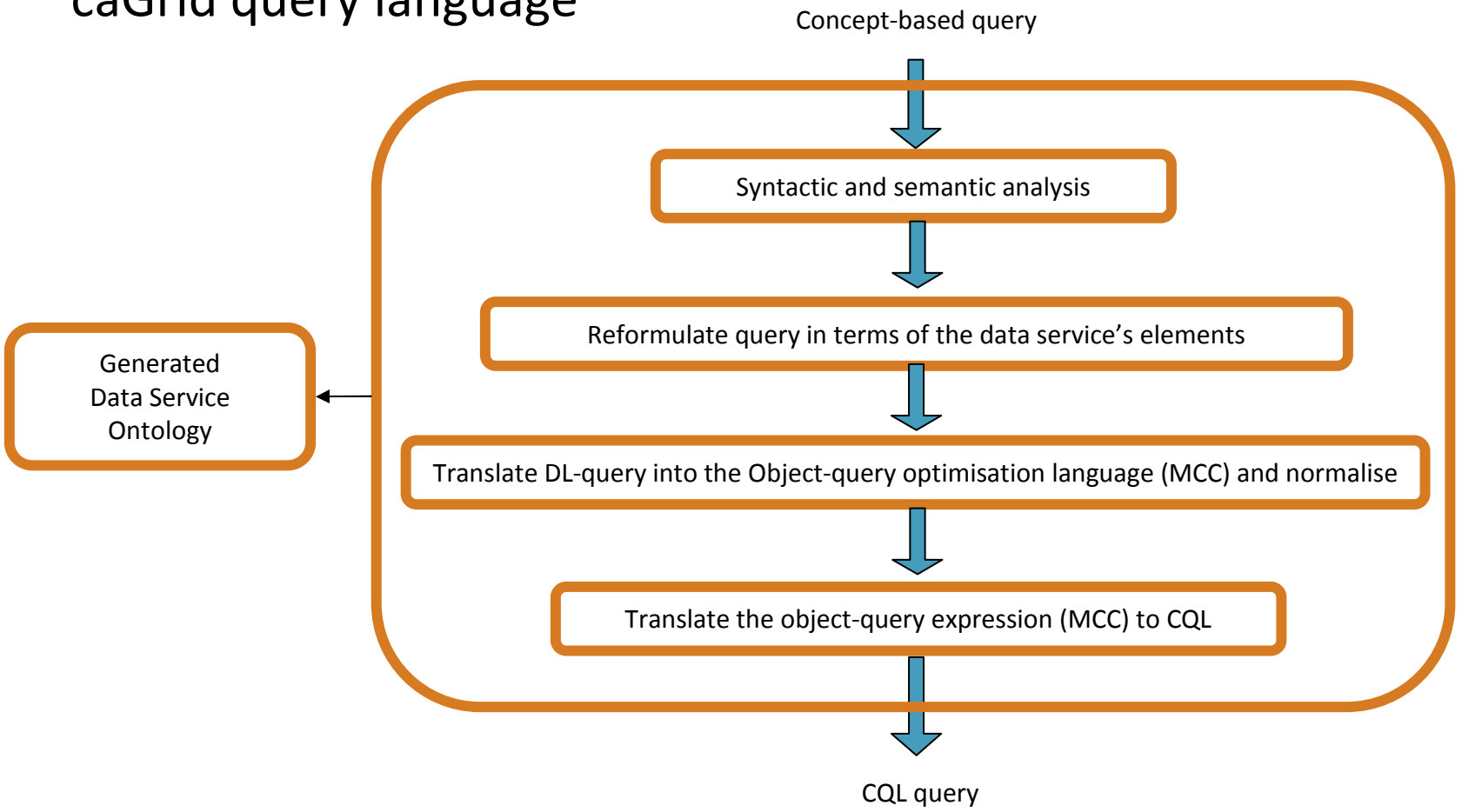
- hasConcept some *Single_Nucleotide_Polymorphism* and hasAssociation some (hasConcept some *Gene* and hasAttribute some (hasConcept some *Gene_Symbol* and hasValue value "*TGFB1*")

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

Approach

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



Conclusions

- Support for high-level descriptive queries based on domain concepts for caGrid data services
- Approach generates ontologies from caGrid metadata and uses ontologies+reasoning to translate concept-based queries to CQL
  - Same concept-based query applicable to all the relevant resources
  - No need for the user to be aware of the structure of the target resource
- General approach: it is applicable to other resources exposing metadata and semantic annotations for query translation and data integration
  - James P. McCusker, Joshua A. Phillips, Alejandra González Beltrán, Anthony Finkelstein, Michael Krauthammer **"Semantic web data warehousing for caGrid"** BMC Bioinformatics SWAT4LS Supplement, in press 2009.
- Prototype Demo **http://tinyurl.com/o6uw7z**

Acknowledgements

- Cancer Research UK
- UK National Cancer Research Institute Informatics Initiative

Thank you!

Questions?