Exposing caGrid Data Services as Linked Data

Joshua Phillips, BS[1] Alejandra González-Beltrán, PhD[2] Anthony Finkelstein, PhD[2] Jyotishman Pathak, PhD[3]
[1]SemanticBits LLC, Herndon, VA [2]University College London, London, UK
[3]Mayo Clinic, Rochester, MN

Abstract. The National Cancer Institute (NCI) enables sharing cancer-related data through an open, federated information network based on the caGrid middleware. The caGrid supports interoperability by building standard-based services with precise semantic definitions. However, rapid, yet flexible, integration of data is not supported. In this research, we address this requirement by exposing the caGrid data services as Linked Data and illustrate the approach integrating data from a tissue bank repository (caTissue) with a microarray/gene expression database (caArray).

Introduction. The NCI's caGrid middleware allows application developers to build data services to share data complying with semantically-annotated UML information models. The semantic annotations correspond to entities from the NCI Thesaurus [1] and provide a rigorous semantics-based interoperable infrastructure. However, the caGrid query facilities fail to provide a transparent mapping of semantics onto the data provided by the data services. Consequently, a caGrid client trying to access data from diverse data services has to perform complex service metadata analysis and mappings to perform federated querying and data integration. The overall objective of the proposed work is to address these limitations by evaluating emerging Semantic Web technologies such as Resource Description Framework (RDF) [2], in the context of Linked Data, for rapid and flexible integration of caGrid data.

Background. Linked Data [8] provides a very powerful framework for heterogeneous data integration. It relies on (i) the simple RDF data model, (ii) de-referenceable Uniform Resource Identifiers (URIs) for creating globally unique names, and (iii) standard languages such as OWL [3], and SPARQL [4] for creating ontologies and modeling and querying data. The Linked Open Data initiative [5] is applying this approach with the aim of bootstrapping the Web of data by publishing existing data sets in RDF and creating numerous links between them. At present, there are more than 70 domain-specific public datasets, with approximately 7 billion triples connected via 140 million links, creating a huge integrated-network dataset on which very expressive queries can be executed.

Methods. The first step of our process involves generating OWL ontologies from the information model (UML model with semantic annotations) for each caGrid data service. The UML-to-OWL transformation (i) uses various semantics-preserving ontology module extraction techniques for selecting relevant NCI Thesaurus entities, (ii) represents the relationships between UML entities, and (iii) models the mappings between UML and NCI Thesaurus entities. The second step comprises translating the actual data from the caGrid data services as RDF that conforms to the ontology generated via the UML-to-OWL transformations. In previous work [6], we generated RDF files by translating caGrid XML (from query results) to RDF triples, which we stored in a data warehouse. While practical and useful, such an approach has limitations in handling large data sets, as well as provides a static view of the data. In this work, we dynamically generated RDF data from caGrid data services using the D2R server [7]. D2R facilitates customizable relational-to-RDF mapping for representing caGrid data as OWL instance data (conforming to the OWL ontologies) that can be queried using SPARQL in a knowledge base. Our approach relies on the caGrid identifier framework to identify caGrid resources with URIs and uses caCORE SDK XMI [9] to modify D2R mappings to use the OWL classes that represent entities from the information models.

Results and Discussion. While our approach can be generalized to integrate data from any caGrid data service, we have developed an initial prototype focused on two existing services: caTissue and caArray. Preliminary evaluation provided insights on the UML-to-OWL transformations as well as on modeling the relational-to-RDF mappings. We have shown that it is useful and practical to expose caGrid data services as Linked Data and SPARQL HTTP interfaces. Our on-going work is focused on improving and extending SPARQL query translations (e.g., SPARQL to HQL).

References

[1] Noy N, de Cornado S, Solbrig H, et al. Representing the NCI thesaurus in OWL DL: Modeling tools that help modeling languages. Applied Ontology 2008. 3(3): 173–190
[2] Resource Description Framework, 2004 http://www.w3.org/RDF/
[3] Web Ontology Language, 2004 http://www.w3.org/2007/OWL/
[4] SPARQL Query Language for RDF, 2008 http://www.w3.org/2009/sparql/wiki
[5] Linked Open Data Initiative, 2009 http://linkeddata.org/
[6] McCusker JP, Phillips JA, González-Beltrán A et al. Semantic web data warehousing for caGrid. BMC Bioinformatics 2009:10.
[7] D2R Server: Publishing Relational Databases on the Semantic Web, 2009 http://www4.wiwiss.fu-berlin.de/bizer/d2r-server/
[8] Berners-Lee, T. Linked Data Design Issues, 2009 http://www.w3.org/DesignIssues/LinkedData.html
[9] Komatsoulis, G. A. et al. caCORE version 3: Implementation of a model driven, service-oriented architecture for semantic interoperability J. of Biomedical Informatics, 2008, 41, 106-123