Building an FHIR Ontology based Data Access Framework with the OHDSI Data Repositories

Guoqian Jiang, MD, PhD[1], Guohui Xiao, PhD[2], Richard C. Kiefer[1], Eric Prod'hommeaux[3], Harold R. Solbrig, MS[1]
[1]Mayo Clinic, Rochester, MN, USA; [2]Free University of Bozen-Bolzano, Italy; [3]MIT/W3C, Cambridge, MA, USA

Introduction

Clinical and translational research increasingly relies on the existence of robust integrated data repositories (IDRs) to combine clinical and "-omics" data. A variety of data models have been developed to provide standardized interfaces to organize research data in a clinical data repository. The Observational Health Data Sciences and Informatics (OHDSI) Common Data Model (CDM) has been increasingly used to build a large-scale international data network in support of observational studies [1]. Other data models include the i2b2 star schema, and the PCORNet CDM. These data models serve well as a layer of standardization for clinical research data within their own research network; however, if investigators want to reuse and integrate applications and accompanying research datasets across different research networks they still face huge challenges. This situation demands a global data model as a reference standard to facilitate data model harmonization and integration.

The HL7 Fast Healthcare Interoperability Resources (FHIR) is emerging as a next generation standards framework for facilitating health care and electronic health records (EHR)-based data exchange [2]. There is a critical need to build the FHIR-based data access and query on existing relational data sources to facilitate standards-based semantic data integration, sharing and discovery in broader scientific research communities [3]. The objective of the study is to develop an FHIR-based data access framework to enable answering semantic queries over the OHDSI CDM-based data repositories. We leverage the FHIR ontology and an open-source Ontology-based Data Access (OBDA) system known as Ontop [4] to demonstrate the feasibility of our approach.

Methods

Figure 1 shows the system architecture, comprising four layers: an input layer, a transformation layer, a semantic query layer and an application layer.

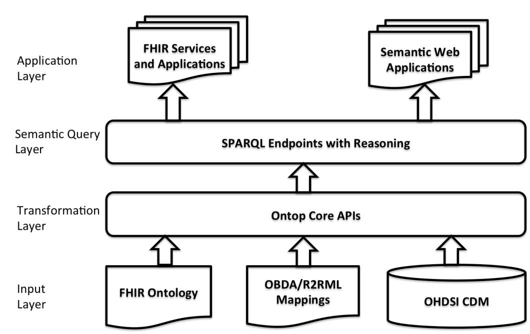

Figure 1. System Architecture

OHDSI CDM and Virtual Machine (VM): The OHDSI CDM 5.0.1 version (https://github.com/OHDSI/CommonDataModel) contains 39 database tables in 6 categories: standardized clinical data, standardized health system data, standardized health economics, standardized metadata, standardized vocabularies and standardized derived elements. We installed an OHDSI VM that is conformant to the OHDSI CDM 5.0.1 version. The VM contains the full OHDSI technology stack and is loaded with both standard vocabularies and sample data. The VM uses a relational database known as PostgreSQL as a storage backend. Ontop and its Protégé 5 Plugin: Ontop is an open-source OBDA system developed for querying relational data sources using an ontology-based approach. We installed the Protégé 5 Ontop Plugin (http://ontop.inf.unibz.it/), comprising 1) the Ontop Mappings tab – managing data source (i.e., database connection), and mapping creation through building an OBDA model (i.e., a specification of how the data in a data source is mapped to the vocabulary in an ontology); and 2) the Ontop SPARQL tab – providing a query editor to allow for editing SPARQL queries and executing a query to test an OBDA model. We at the Mayo Clinic have been collaborating with the FHIR and W3C HCLS community to develop the FHIR Resource Description Framework (RDF) representation specification and associated transformation and validation tools [2]. One of such efforts is to produce the FHIR *StructureDefinition* resource to OWL transformation, known as the "FHIR Ontology" (http://build.fhir.org/fhir.ttl). The *StructureDefinition* resource is the metamodel for FHIR resource definitions, meaning that a FHIR resource such as *Patient*, is formally defined using an instance of *StructureDefinition* that declares elements like "*Patient.name*" and "*Patient.birthDate*" and associated metadata and constraints (e.g., datatype and cardinality). The FHIR Ontology formally enumerates the classes, predicates, domains, ranges and specific datatypes that are used in describing the FHIR instance data.

We first loaded the FHIR Ontology into a Protégé 5 environment, and used the Datasource manager in the Ontop Mappings tab and established the database connection to the OHDSI VM. And then we used the Mapping manager and Mapping assistant in the same tab and created an OBDA model declaring mappings between the FHIR Ontology and the OHDSI CDM database schema. The consensus of the mappings was achieved through a group discussion among co-authors. We also extended the FHIR Ontology with a number of datatype properties to capture the values of primitive datatypes. Once the datasource is connected and the mappings are defined, the Ontop Core APIs can be invoked and the SPARQL endpoints with reasoning capability (in terms of OWL 2 QL and RDFS) can be established. In this study, we used the Ontop SPARQL plugin that encapsulates both transformation and semantic query layers for the testing purpose. We created a collection of the SPARQL queries representing clinical research data questions, and used the queries to test the system.

Results and Discussion

We created the mappings for 7 OHDSI CDM tables: *Person, Condition Occurrence, Drug Exposure, Observation*, *Concept, Concept Relationship* and *Concept Ancestor*. Out of 7 tables, 5 tables are mapped to the FHIR core resources, i.e., the Person table to *fhir:Patient* (with 2 properties), the Condition Occurrence table to *fhir:Condition* (5 properties), the Drug Exposure table to *fhir:MedicationAdministration* (6 properties), the Observation table to *fhir:Observation* (7 properties). The rest of tables are from the Vocabulary CDM. The Concept table is mapped to a complex FHIR datatype Coding (3 properties), and the Concept Relationship table and the Concept Ancestor table are mapped to the FHIR resource ConceptMap (3 properties). A total of 418,602,850 virtual triples were produced through the mappings. We successfully created the SPARQL query templates that can identify patient cohorts using a single concept code or its descendants (i.e., demonstrating inference capability) from different domains (i.e., Condition, Medication Administration, and Observation). **Table 1** shows the query results of three example concept codes in three domains. We verified that the results are accurate using an OHDSI open source cohort identification tool known as ATLAS (https://github.com/OHDSI/Atlas).

Table 1. Query results of three example concept codes in three different domains.

Domain	Concept Code	Descendants	Number of Patients Retrieved with a Single Code	Number of Patients Retrieved with Descendants
Condition	Acute myocardial infarction|57054005|SNOMED	57	6757	14453
Medication Administration	Warfarin|11289| RxNorm	489	0	25602
Observation	past history of procedure|416940007| SNOMED	481	4678	56772

Discussion

In this study, we developed an ontology-based data access framework for enabling FHIR-compliant semantic queries over the OHDSI CDM-based data repositories. We have demonstrated the feasibility of the framework by implementing a prototype leveraging the FHIR ontology and an open-source Ontop system. Our ongoing next steps include: 1) building a set of robust mappings between FHIR and OHDSI CDM through a community-based consensus approach; 2) enhancing the Ontop OBDA model to handle blank nodes which are required for the FHIR canonical RDF representation; 3) producing a library of SPARQL query templates for complex cohort identification logic; and 4) enabling federated queries across the OHDSI data repositories and beyond.

Acknowledgement

This study is supported in part by NIH grants U01 HG009450, U01 CA180940, and R01 GM105688.

References

1. OHDSI URL: https://www.ohdsi.org/; last visited at March 6, 2017.
2. FHIR. http://hl7.org/fhir; last visited at March 6, 2017
3. Solbrig HR, Prud'hommeaux E, Grieve G, McKenzie L, Mandel JC, Sharma DK, Jiang G. Modeling and Validating HL7 FHIR Profiles Using Semantic Web Shape Expressions (ShEx). J Biomed Inform. 2017 Feb 14. pii: S1532-0464(17)30034-5.
4. Calvanese D, Cogrel B, Komla-Ebri S, Kontchakov R, Lanti D, Rezk M, Rodriguez-Muro M, and Xiao G. Ontop: Answering SPARQL queries over relational databases. Semantic Web, vol. 8, no. 3, pp. 471-487, 2017.