# Using Semantic Queries for Cohort Discovery Across Research Networks

**Amanda Hicks, PhD[1], William R. Hogan, MD, MS[1], Zhe He, PhD, MS[2], Josh Hanna, MS[1], Betsy Shenkman, PhD, MSN[1], Jiawei Yuan, PhD[3], and Jiang Bian, PhD, MS[1]**

**[1]Univsesity of Florida, Gainesville, Florida; [2]Florida State University, Lake City, Florida; [3]Embry-Riddle Aeronautical University, Daytona Beach, Florida**

**Abstract**

*Datasets from research networks are rapidly growing in both number and variety. This raises the question: how can we best integrate heterogeneous datasets? Even with common data models (CDM) for each research network, we still face the problem of how to query across networks using different data models. We propose a framework design that uses Semantic Web technology to query across different CDMs and demonstrate its potential using realistic use cases for cohort discovery.*

**Introduction**

The last few years have witnessed an increasing number of research networks including the National Patient-Centered Clinical Research Network (PCORnet) and Accrual to Clinical Trials (ACT). These networks target building regional and national collections of data from electronic health records (EHRs), claims, and patient-reported outcomes (PROs). Datasets are growing rapidly, not just in volume, but also in variety, which presents challenges in integrating datasets from heterogeneous sources. Even when data are transformed and integrated into a structured format using a common data model (CDM) from raw datasets for a research network, we still face the problem of how to query across networks using different data models.

Reconciling different data models and transferring data from different networks into another CDM is not cost-effective. For example, aligning data elements in different data models to create a superset CDM costs time and effort. It will also be difficult to keep the superset CDM up-to-date when individual CDMs change. Further, this method creates another set of Extract-Transform-Load (ETL) processes, which leads to data and information quality issues. Last but not the least, replicating datasets into different formats not only incurs additional costs in computing recourses (e.g., storage) and personnel, but more importantly increases security risks.

These observations motivated us to investigate strategies for querying across research data networks that use different CDMs. Even though many CDMs leverage ontologies and standard terminologies to give the data model some semantics (e.g., CTSA ACT Ontology and i2b2 Ontology), very little work has focused on using Semantic Web technology (e.g., ontology and semantic query) to query across different CDMs. In this work, we propose a design of such framework and demonstrate its potential using realistic use cases for cohort discovery. In essence, we align data elements in each individual CDM to existing biomedical ontologies, develop a cohort discovery ontology (CDO) framework reusing those ontologies as modules, and include domain specific classes and relations where necessary for individual studies as study specific application ontologies (see Figure 1 in the poster). We then use semantic queries (e.g., using the SPARQL Protocol and RDF Query Language) instead of conventional SQL (Structured Query Language) queries. A number of efforts[1] have been focused on providing SPARQL queries over relational databases, which most research networks' data infrastructure is relied upon. SPARQL queries support inference over ontologically modeled data using inheritance relations and logical definitions for defined classes. For example, if we are looking for patient populations with Phelan-McDermind syndrome, some patients may not have a diagnosis yet still exhibit symptoms or have a diagnosis of autism spectrum disorder or intellectual disability. A SPARQL query that uses application ontology that models these signs, symptoms, and comorbidities of Phenlan-McDermind syndrome can be used to retrieve a count or list of patient IDs based on a query for the disease rather than creating a full list of possible diagnoses and attributes for each CDM in a SQL query.

**Methods**

The first step in our proposed framework is to select a suite of ontologies to use for developing the CDO framework and the SPARQL queries. The CDMs are then mapped to the ontologies in the CDO framework. These ontologies can be used for simple queries, but application ontologies that are specific to a domain such as a disease or a particular set of drug-drug interactions may produce more accurate results. For such cases, an application ontology that models domain specific classes and relations are integrated into the CDO framework to develop a set of domain specific query ontologies. Finally SPARQL queries are constructed for cohort identification.

For this paper, we have modeled demographic data from OMOP and PCORnet using the Ontology of Medically Related Social Entities (OMRSE)[1]. OMRSE an ontology that is developed in accordance with the OBO Foundry's best practices and reuses classes from many other biomedical ontologies. It also models some demographic

information and social history, in particular, gender, race, ethnicity, and smoking status. Next we analyzed the classes required to find cohort using demographic data stored in Observational Medical Outcomes Partnership (OMOP) CDM and the PCORnet CDM, adding new classes to OMRSE where necessary. Finally, we constructed SPARQL queries that utilize this demographic data and that meet the use cases from PCORNet data harmonization efforts discussed in the January 2014 Workshop for Data Harmonization for Patient-Centered Clinical Research.[2] We focused on race, ethnicity, gender, smoking status and period of complete data capture for this stage of the development.

**Results**

Most CDMs follow best practices for interoperability to use standard terminologies as much as possible, which makes the modeling process easier. Because OMOP and PCORnet frequently use the same coding scheme, the comparison of the values for demographic data is straight forward. For example, both CDMs us HL7's administrative gender codes as values for sex/gender. Similarly, OMOP uses the CDCs classification and PCORnet use the standard created by the Office of Management and Budget (OMB) to code race. However, the CDC's classification of race is explicitly an elaboration on the OMB's classification and has been mapped to it already[5].

We identified three use cases that fit three of the four use case types described in the 2014 Workshop on Data harmonization for Patient-Centered Clincial Research[6] that utilize the demographic information that we have represented. Our use cases are 1) count unique individuals who are smokers, 2) count the number of smokers in the network by race, and 3) retrieve a list of patients who are black or African American and are smokers.

We then wrote SPARQL queries for each use case and verified them over small knowledge base. For example, we constructed the following query for the first use case:

> *SELECT DISTINCT ?personCount (COUNT(?person) as ?personCount) WHERE { ?role rel:inheres_in ?person . ?role a obo:OMRSE_00000039 . } GROUP BY ?person*

**Discussion**

In this poster we focus on demographic information. However, the approach can be extended to other types of patient data by leveraging other ontologies that are developed according to OBO best practices. For example, the Drug Ontology[3] can be used to find patients who are prescribed drugs described by RxNorm ID, class of drug, indication, active ingredient, or mechanism of action. Likewise, the Human Disease Ontology[4] contains mappings to SNOMED-CT codes and can be used for cohort identification by diagnosis code.

We currently have a working group that focuses on ontologically representing the PCORnet CDM in OBO-compliant ontologies. In addition to this work, future steps for implementing this framework include ensuring that other CDMs are fully represented in the ontologies. Further work also needs to be done on mapping CDMs to a set of ontologies in the CDO framework, from which one can develop an application ontology for domain specific cohort identification queries.

**Conclusion**

Cohort identification can be achieved across research networks with different CDMs using SPARQL queries with an application ontology, mappings from CDMs to the ontology, and a SPARQL support service over existing relational databases.

**Acknowledgement**

This work supported in part by the NIH/NCATS Clinical and Translational Science Awards to the University of Florida UL1 TR000064 and the OneFlorida Clinical Research Consortium.

**References**

1. Hogan WR, Garimalla S, Tariq SA. Representing the reality underlying demographic data. ICBO; 2011.
2. Institute of Medicine. Data harmonization for patient-centered clincial research - a workshop 2014 [cited 2015 August 30]. Available from: http://iom.nationalacademies.org/~/media/Files/Activity
3. Hogan, WR., Hanna J, Joseph E, Brochhausen. Towards a consistent and scientifically accurate drug ontology. ICBO; 2013.
4. Disease ontology [cited 2015 August 30]. Available from: http://do-wiki.nubic.northwestern.edu/do-wiki/index.php/Main_Page.
5. McFadden, B, Nerenz DR, and Ulmer C, editors. Race, Ethnicity, and Language Data: Standardization for Health Care Quality Improvement. National Academies Press, 2009.
6. Spanos DE, Stavrou P, Mitrou N. Bringing relational databases into the Semantic Web: A survey. *Semantic Web*. 2012:*3*(2);169-209.