Using Ontologies to Identify Patients with Diabetes in Electronic Health Records

Hairong Yu, Siaw-Teng Liaw, Jane Taggart, and Alireza Rahimi Khorzoughi

School of Public Health & Community Medicine and Research Centre for Primary Health Care & Equity, Faculty of Medicine, University of New South Wales, Sydney, NSW 2052, Australia

{hairong.yu,siaw,j.taggart}@unsw.edu.au
alireza.rahimikhorzoughi@student.unsw.edu.au

Abstract. This paper describes a work in progress that explores the applicability of ontologies to solve problems in the medical domain. We investigate whether it is feasible to use ontologies and ontology-based data access (OBDA) to automate common clinical tasks faced by general practitioners (GPs), which are labor-intensive and error prone in terms of relevant information retrieved from electronic health records (EHRs). Our study aims to improve the selection of diabetes patients for clinical trials or medical research. The biggest impediment to automating such clinical tasks is the essential bridging of the semantic gaps between existing patient data in EHRs, such as reasons for visit, chronic conditions and diagnoses, pathology tests and prescriptions stored in general practice EHRs (GPEHR), and the ways which medical researchers or GPs interpret those records. Our current understanding is that automated identification of diabetes patients can be specified systematically as a solution supported by semantic retrieval. We detail the challenges to building a realistic case study, which consists of solving issues related to conceptualization of data and domain context, integration of different datasets, ontology creation based on the SNOMED CT-AU® standard, mapping between existing data and ontology, and the challenge of data fitness for research use. Our prototype is based on data which scale to thirteen years of approximately 100,000 anonymous patient records from four general practices in south western Sydney.

Keywords: Ontology, Diabetes Mellitus, Electronic Health Records, eHealth, Knowledgebase Management, Ontology_Based Database Access

1 Introduction

This paper reports on work that explores the applicability of ontologies for solutions in the health domain. In Australia, the main health applications of ontologies appear to be the SNOMED terminology services. We investigated the feasibility of the use of ontologies and OBDA to automate clinical tasks, such as identifying patients with specific diabetes mellitus (DM) phenotypes in EHRs contributing to the data repository of the electronic Practice Based Research Network (ePBRN). The ePBRN is used

to conduct translational research on primary and integrated care, including tracking patients, managing chronic disease, and providing quality evidence-based care. The biggest barrier to automating the clinical task of identifying a patient with DM is the semantic gaps between patient data in EHRs, such as reasons for visit, diagnoses, pathology tests and prescriptions, and how these EHR data are interpreted. For example, in addition to a diagnostic label, DM can be implied by a blood glucose test with suggestive levels of diabetes, certain medications such as oral hypoglycaemics or insulin, or the use of DM supplies such as glucose diagnostic strips. By using ontologies, our experiments show that it is possible to automate this interpretation process and build a reusable conceptual infrastructure over diverse standards or experience or datasets. Currently most efforts at automation is only limited within individual clinics or in a physician-driven process or at data levels.

The SNOMED CT-AU®, the Australian extension to SNOMED CT® (Systematized Nomenclature Of Medicine Clinical Terms), is an ontology which formally defines classes of medical procedure, pharmaceutical or biologic product, and body structure and so on. The SNOMED CT-AU® Ontology (SCAO) is the reference terminology for EHRs in Australia. SCAO is available in Web Ontology Language (OWL) format from the Australian National E-Health Transition Authority (NEHTA). Our experiments showed that the integration of SNOMED CT-AU and the Diabetes Identification Ontology (DIO) based on ePBRN data to select patients with DM is well suited for our case study. Our key approach is that the automation of the process of identifying DM patients is an issue of semantic retrieval, i.e. selection criteria can be expressed as semantic queries, which are processed by a reasoner to retrieve explicit information on eligible patients from datasets and infer implicit knowledge from ontologies simultaneously.

The objective of this study is to assess the practicality and utility of ontologies in a real world environment. The technical challenges of conceptualization of data and domain context, ontology integration of different datasets or ontologies, mapping between existing datasets and ontologies, and finding solutions to ensure data fitness for clinical or research use will be described and discussed in the following sections.

2 Methodology

The architecture for this study comprises six parts separated by dashed lines as shown in Figure 1. Patient data were extracted from individual GPEHRs, e.g. Medical Director™[1] at each clinic by GRHANITE™[2]. The software provides a data repository over server called GRHANITE™ Databank, in our case the ePBRN repository operated by MS SQL Server™. The ABox, associated with instances of ontology classes or properties, is populated through ontopPro (formerly known as an OBDA plugin for Protégé[3]). Another primary component in our knowledgebase, the TBox, related to concep-

[1] http://www.hcn.com.au/Products/Medical+Director
[2] http://www.grhanite.com/
[3] http://protege.stanford.edu/

tual terminologies defined in ontologies, is built through Protégé, a popular open source ontology editor and knowledgebase framework.

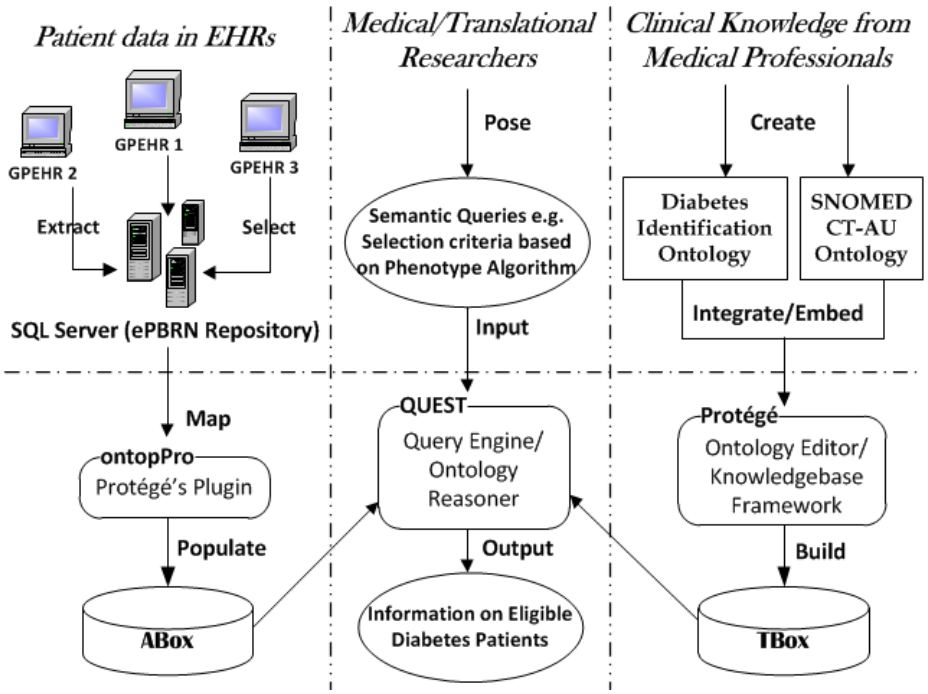

Fig. 1. ePBRN Diabetes Identification Case Study Solution Architecture

Clinical selection criteria are formulated as semantic queries in SPARQL Protocol and RDF Query Language (SPARQL). The SPARQL query engine QUEST[4] that comes with ontopPro[5] checks the queries against the knowledgebase to retrieve matched patients. A demonstration is given for each of six parts in Figure 1.

The first step in building this solution is creating the specific DIO in hierarchical conceptual modeling, based on the Australian National Guidelines for Type 2 Diabetes Mellitus (T2DM) and discussions with the research team and GPs participating in the ePBRN. The output of this first task is a formalized ontology which consists of 4 main classes Actor, Content, Mechanism and Impact and 68 subclasses with object/data properties. Some of them can be mapped to the SNOMED CT-AU Ontology (SCAO), which has more than 300,000 concepts.

Due to the small number of concepts captured by the DIO, the mapping can be operated manually. For example, T2DM is a Disease under the subclass of Problem which has a superclass Context in DIO. In the SCAO, T2DM is a disorder of glucose metabolism which is a subclass of Disease under the highest level concept of Clinical

[4] http://semanticweb.org/wiki/Quest/
[5] http://ontop.inf.unibz.it/

finding. Similarly, Actor class in DIO corresponds to Environment or Geographical location in SCAO. However the automation of integration of two ontologies can be complex for large terminologies.

Next we linked the server objects in SQL Server to integrate other heterogeneous datasets by T-SQL™. The SQL query results are mapped by ontopPro for ABox associated with relevant classes in ontologies. This meant that the schematic or semantic heterogeneity challenges faced were solved at either data or ontology level. The mapping mechanism supplied by ontopPro theoretically based on OBDA [1], provided a big advantage on populating class members, assigning property values, and incorporating schematic data in the ePBRN repository with semantic concepts in ontologies. The raw data in EHRs that contribute to the ePBRN repository are incomplete, incorrect and inconsistent (against external standards or internal logic perspectives). We used definitions of properties in DIO or mappings created in ontopPro to solve core data quality issues before preparation of semantic queries.

We then wrote semantic queries in SPARQL according to requirements from domain experts, and ran them through QUEST, the query engine and OWL reasoner. The query results are expected to identify DM patients and help clinicians to manage the cycle of care for the cohort. The SPARQL queries were validated using SQL over an artificial dataset of 100 patients schematically similar to the ePBRN dataset. The approach that we developed and tested on the artificial dataset will be scalable to the ePBRN repository of more than 100,000 patient records. Other use case scenarios, for example assisting researchers to conduct association and/or controlled studies will contribute to the validation of the architecture.

3 Discussion and Conclusion

We have briefly presented a feasibility study of the use of ontologies to detect patients with DM in real world EHRs. Using real patient datasets, we solved some engineering challenges around ontology creation and integration, bridging between ontologies and datasets, and data quality [2]. Apart from usability, interoperability and scalability aforementioned, other quality attributers are assessed closely for architecture evaluation for instance, modifiability with many facades/locations where data/data types are transferred in our solution, integrability and extensibility which are especially critical as several open source software components are used in our design.

References

1. M. Lenzerini, Data Integration: A Theoretical Perspective, Proc. of the 21st ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (PODS'02), pp. 233 – 246.
2. S.-T. Liaw, et al. Data quality and fitness for purpose of routinely collected data – a general practice case study from an electronic Practice-Based Research Network, in: AMIA Annual Symposium Proceedings, 2011:785–794.