User Centred Networked Health Care
A. Moen et al. (Eds.)
IOS Press, 2011

doi:10.3233/978-1-60750-806-9-729

Large Scale Healthcare Data Integration and Analysis using the Semantic Web

John TIMM[a], Sondra RENLY[a], Ariel FARKASH[b]
[a] *IBM Almaden Research Center 640 Harry Rd, San Jose, CA, 95120, US*
[b] *IBM Haifa Research Lab, Haifa Univ. Mount Carmel Haifa, 31905, Israel*

Abstract. Healthcare data interoperability can only be achieved when the semantics of the content is well defined and consistently implemented across heterogeneous data sources. Achieving these objectives of interoperability requires the collaboration of experts from several domains. This paper describes tooling that integrates Semantic Web technologies with common tools to facilitate cross-domain collaborative development for the purposes of data interoperability. Our approach is divided into stages of data harmonization and representation, model transformation, and instance generation. We applied our approach on Hypergenes, an EU funded project, where we use our method to the Essential Hypertension disease model using a CDA template. Our domain expert partners include clinical providers, clinical domain researchers, healthcare information technology experts, and a variety of clinical data consumers. We show that bringing Semantic Web technologies into the healthcare interoperability toolkit increases opportunities for beneficial collaboration thus improving patient care and clinical research outcomes.

Keywords. Healthcare Interoperability, Semantic Web, Modeling, UML, OWL.

1. Introduction

Healthcare providers want access to healthcare information to improve coordination of care, increase quality of care, and generate evidence for future medical decision making. In Hypergenes project [1], we aimed to integrate and analyze heterogeneous hypertension data sets from over 30 historical cohorts spanning 15 years with the goal of creating a new data model for representing interactions between environmental and clinical factors in hypertension. A successful outcome will lead to improved diagnostic accuracy, early detection, and personalized treatments. In the past, these types of research efforts involved the development of a new data sharing infrastructure which is time consuming and costly; thus preventing large scale data integration and analysis. Today Semantic Web technologies, new tooling, and healthcare data standards enable this type of large scale integration and analysis. Building on the world-wide web's scalable, distributed architecture for sharing information efficiently between humans, the Semantic Web provides capabilities that enable information sharing between machines that are semantically consistent. The Semantic Web consists of a set of standards and technologies that include a simple data model (RDF), query language (SPARQL), schema language (RDFS) and ontology language (OWL). These technologies assist in data integration from heterogeneous data sets. Healthcare providers are moving towards health information standards for sharing subsets of patient records using specialty-developed Implementation Guides (IGs) built using

standards such as HL7 v3 Clinical Document Architecture (CDA) and aligned with the CEN EHR 13606 specification. Clinicians use the shared content to make more informed medical decisions for their patients and to better coordinate care when their patients get care from multiple sources. In this paper, we depict a methodology that aims to improve data integration, analysis and sharing between clinical information systems and researchers. Our approach brings together standard healthcare information models with semantic web technology in an effort to accommodate multiple user roles and leverage the strengths of different technologies to address specific aspects of the healthcare interoperability problem.

2. Background & Related Work

Biomedical information repositories typically contain data related to a specific clinical domain with proprietary semantics [2]. These disparate data sources pose a challenge for data integration [3] that is paramount for improved patient-centric care [4], health data exchange, decision support [5], and semantic query and retrieval of aggregated data for analysis in context of clinical research. CDA is a health information standard that specifies terminology-encoded structure and semantics for clinical documents. CDA documents can be serialized to XML that conforms to a published W3C XML Schema. In most applications, the general CDA structure is constrained by a set of templates that are standardized and published in an implementation guide, such as the Continuity of Care Document (CCD). As in most CDA template specifications CCD IG is written in structured English expressions based on the XML schema element relationships. These conformance statements are usually implemented by Schematron rules to augment the CDA XML schema. Our work includes methods and open source software tools for representing CDA documents and template constraints using Unified Modeling Language (UML) and Object Constraint Language (OCL).

The UML modeling language is dominant among IT domain users, whereas clinical domain experts often work with formal ontology definitions. Web Ontology Language (OWL) is a semantic markup language for publishing and sharing ontologies on the World Wide Web. It is endorsed by the World Wide Web Consortium (W3C). OWL is often used as the framework for converging distinctive terminologies into a single coherent ontology; many successful examples exist in clinical research and medical informatics domains [6,7]. For ontology mapping we followed W3C recommendations. There has been some prior work in both using OWL ontologies in conjunction with instance generation [8], and in using OWL to add semantic annotations to UML information models [9]. We extended these to support our multi-faceted approach.

3. Methods

Our solution (figure 1) starts with a clinical domain researcher (upper left) creating an ontological representation of the information elements of interest needed for a particular study known as a *cohort ontology*. Ontologies from all data sources are mapped to a common *core ontology*. Based on past experiences, the clinical domain expert is less interested in the comprehensive data representation than in certain data elements in their proper context that are required for further analysis. A leading design

principle of our methodology is to have the clinical domain expert work with an "intuitive" ontology-based method to represent the metadata needed for harmonization, while the healthcare IT domain expert uses modeling languages and semantic web technologies to create, constrain and transform representations of the standard format. The point of collaboration is focused at mapping core ontology to data representation creating a warehouse that is standard, interoperable and allows for semantic query and retrieval of data in research oriented scenarios.

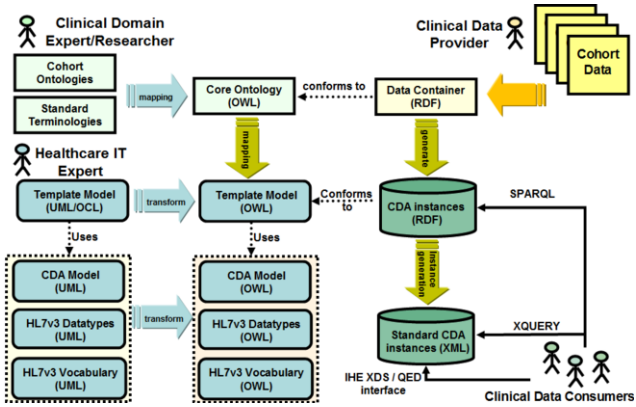

Figure 1. Data Integration Methodology Overview.

The healthcare IT expert (lower left), familiar with data representation methods and standards, is primarily responsible for creating healthcare interoperability models. These models will be used to derive the common format that is collected and subsequently analyzed. We use models based on international standards for healthcare semantics and interoperability that can be serialized to XML. These, along with a set of constraints, serve to unify data into a semantically unambiguous format that makes operations on the data straightforward from a technological standpoint.

Integration of data from dissimilar data sources including harmonization, data extraction, validation and normalization is a complex task due to ambiguous metadata, differences in units of measurement, classifications, diversity of protocols, etc.; the process is described at length in previous works [8,10]. Thus in the clinical research scenario we will assume the clinical data provider (upper right) supplies RDF that conforms to the cohort ontology, then by using the cohort to core mapping, the data graph is converted to conform to the core ontology. Mappings between the ontological representation and semantic data representation enable generation of RDF instances that conform to healthcare interoperability models which in turn are fed to an instance generation engine in order to produce standard XML instances. In the health-oriented scenarios standardized data is received either via IHE XDS source or directly inserted using a simple adapter into the XML database. The CDA instance received conforms to the template model; using the UML to OWL model transformation we convert it to an RDF instance that conforms to the OWL template model. Clinical data consumers may then access data via interoperability profiles, e.g. IHE XDS/QED, query XML database directly using XQuery, or query data semantics using SPARQL.

The CDA UML model was created as an implementation model that is primarily based on two artifacts: (1) the CDA Refined Message Information Model from HL7 and (2) the CDA XML Schema. This implementation model was developed to support

the existing code generation and serialization mechanisms present in the Eclipse Modeling Framework (EMF). The model was imported into EMF and ultimately transformed into a set of Java classes as a part of the Model-Driven Health Tools (MDHT) [11] project in Open Health Tools (OHT). The Java classes in conjunction with a set of additional utility classes make up the base runtime API that can be used to produce, consume, and validate instances of CDA. The template model is a domain-specific model that constrains the CDA model. Classes in a template model extend those in the CDA model. Constraints are modeled using directed associations, property redefinitions, and OCL expressions. The CDA Profile for UML is used to capture additional metadata needed during model transformation and at runtime. Once the template model has been created, it is transformed into an implementation model which leads to the generation of a domain-specific API for constructing and validating instances. All directed associations, property redefinitions and metadata specified in the template model are converted to OCL expressions in the implementation model.

Leveraging technology from the Semantic Web enables the transformation of the data models to an OWL representation. Many of the constraints can be modeled using OWL restrictions. For example, a fixed or default value in the template model is translated to an OWL value restriction and a directed association is translated to an OWL cardinality restriction. Some constraints, specified in general OCL expressions, are not readily converted into OWL restrictions. Part of our ongoing research is to determine the best mechanism to represent these types of constraints. Possibilities include a semantic rule language such as the Semantic Web Rule Language (SWRL) or Jena Rules. Connecting the core ontology created by the clinical domain expert and the template model a product of the Healthcare IT expert is a crucial step that requires their collaboration. Core ontology variables and their possible parameterizations are mapped via equivalent class and equivalent property relationships to the template model ontology using Jena API following OWL mapping W3C recommendation.

4. Results & Discussion

The Hypergenes project, a Seventh Framework Program (FP7) European Commission funded project exploring the EH disease model, provided us with an opportunity to apply our approach to widely varying environmental and clinical datasets from over thirty historical cohorts. The data included historical clinical data spanning over 15 years and environmental measures based on questionnaires for a total of 8,000 subjects divided into a discovery phase (4000) and a validation phase (4000).

The first phase of the project was aimed at defining the corresponding terminology so that all the cohorts' variables could be mapped to a uniform terminology. Domain experts from the data sources helped define a core hypertension ontology. We then mapped each cohort metadata to this uniform core ontology. The next step involved capturing data semantics using the core ontology. To this end we created the CDA based Essential Hypertension template model (EH-CDA). The quantity, diversity and complexity of data in Hypergenes forced a situation where there was a need to create a large number of templates. This made the modeling process time consuming, challenging, and error prone. Thus we used an automated approach to generate the UML template model from a prototypical XML instance [12]. An OWL representation of the EH-CDA model was then generated from the UML representation. Each template in the model, represented as a UML class specializing the base CDA model,

was converted to an OWL class with a *subClassOf* relationship to the corresponding class in the CDA ontology. Mapping between the path of the UML class in the template model to the generated OWL class was captured and used by the instance generation engine to produce standard XML instances from RDF triples that conform to the model ontology.

We developed several mechanisms for accessing the data. Three are described in the methods section: SPARQL endpoint for semantic querying, direct access to the database via XQuery, and data access via standard IHE XDS and QED profiles. However, the partners in charge of analysis in the Hypergenes consortium use tools that rely on a relational schema. For this purpose we built an RDF to relational module. To accomplish this we built an RDFS that represents the relational schema requested by the analytics partner, and wrote an automatic process that creates the relational schema and populates it. For the population we rely on the SPARQL endpoint, thus, we run a query and insert the result into the corresponding tables in the relational schema.

5. Conclusion

Increasing requirements to implement IG based data exchanges has highlighted the need for expert tailored tooling, established shared core ontologies, mapping processes, and validation technologies. We describe a methodology that improves data integration, analysis, and sharing between clinical and information systems and researchers. We incorporate domain specific user intuitive tools through the transformation path and applied it in Hypergenes EU project. We believe that semantic data instance generation based on standard information models and terminologies serves as a common language that can improve patient care and clinical research outcomes.

References

[1] EC FP7 Hypergenes, http://www.hypergenes.eu/
[2] Stroetmann, V. et al. Semantic Interoperability for Better Health and Safer Healthcare. S*emanticHEALTH Project Report*. http://ec.europa.eu/information_society/ehealth
[3] Heiler S. 1995. Semantic interoperability. *ACM Computing Surveys* 27(2), 271-273.
[4] Gold J. D., Ball M. J. 2007. The Health Record Banking imperative. *IBM Systems Journal* 46(1).
[5] Bock B.J. et al. 2003. The Data Warehouse as a Foundation for Population-Based Reference Intervals. *American Journal of Clinical Pathology* 120, 662-670.
[6] Schultz S., Boeker M., Stenzhorn H. 2008. How Granularity Issues Concern Biomedical Ontology Integration. *MIE*, 863.
[7] Golbreich C., Zhang S., Bodenreider O. 2006. The foundational model of anatomy in OWL: Experience and perspectives. *Web Semantics: Science, Services and Agents on World Wide Web* 4(3), 181-195.
[8] Farkash A. et al. 2006. Biomedical data integration - capturing similarities while preserving disparities. In *Conf Proc IEEE Eng Med Biol. Soc.* 2006 1, 4654-4657.
[9] Carlson, D. 2006. Semantic Models for XML Schema with UML Tooling, In *Proceedings of SWESE 2006*.
[10] Carlson, D. et al. A Model-Driven Approach for Biomedical Data Integration. In *Proceedings of MEDINFO 2010*.
[11] Model-Driven Health Tools (MDHT), http://mdht.projects.openhealthtools.org
[12] Farkash, A. 2010. Facilitating the creation of semantic health information models from XML contents. In *Proceedings of CSHALS 2010*.