Towards Semantic-Web Based Representation and Harmonization of Standard Meta-data Models for Clinical Studies

Cui Tao*, PhD, Guoqian Jiang*, PhD, Weiqi Wei, MM, Harold R. Solbrig, and Christopher G. Chute MD, DrPH

Division of Biomedical Statistics and Informatics, Mayo Clinic College of Medicine, Rochester, MN

Abstract

In this paper, we introduce our case studies for representing clinical study meta-data models such as the HL7 Detailed Clinical Models (DCMs) and the ISO11179 model in a framework that is based on the Semantic-Web technology. We consider such a harmonization would provide computable semantics of the models, thus facilitate the model reuse, model harmonization and data integration.[1]

Introduction

The use of common standardized information building blocks for data capturing and reporting facilitates the understanding and sharing of clinical research information. For instance, the National Cancer Institute (NCI) supports a broad initiative to standardize the Common Data Elements (CDEs) used in cancer research data capturing and reporting [1]. Notably, NCI developed the Cancer Data Standards Repository (caDSR) and chose the ISO/IEC 11179 Metadata Registry standard for metadata registries to represent the CDEs in the database, and implemented a set of APIs and tools used to create, edit, control, deploy and find the CDEs for metadata consumers and for UML model development [1-2].

To build a standard on Detailed Clinical Models (DCMs) is another instance of such an effort by HL7. A DCM is defined as an information model of a discrete set of precise clinical knowledge which can be used in a variety of contexts [3-4]. The DCMs are the refinement of Domain Analysis Models (DAMs), which are in turn the refinement of the HL7 Reference Information Model (RIM). The purpose of DCMs are to provide precise semantic consistent data and terminology specification that are comparable and sharable between multiple care providers, health enterprises and standards-based Healthcare Information Technology (HIT).

While the interactions between information model and terminology are central to achieving practical data standardization, however, the challenges on harmonization of multiple information models and sophisticated terminology models are non-trivial as there is no single unified information model to support clinical research needs [5-6].

Formal knowledge models and knowledge-based methods can be useful on dealing with the challenges. Description Logics (DLs) are a class of knowledge-representation formalisms that are used to represent the terminological knowledge of an application domain in a structured way. The most notable success so far is the adoption of the DL-based Web Ontology Language (OWL) [7] as the standard ontology language for the Semantic Web. OWL was developed for ontology modeling by building hierarchies of classes describing concepts in a domain and relating the classes to each other using properties. OWL can also represent data as instances of OWL classes - referred to as individuals - and it provides mechanisms for reasoning with the data and manipulating it. Rector et al. developed OWL-based methods for defining a Code Binding Interface which have been used in a successful test of the binding of HL7 messages to SNOMED-CT codes [8]. Some efforts have been taken to investigate a model that enables reuse of common observation models across the clinical trials and clinical practice contexts, and how semantic web specification such as OWL can be leveraged [9,10].

In this study, our hypothesis is that representing clinical study meta-data models like HL7 DCMs and the ISO11179 model in a Semantic-Web based framework would provide computable semantics of the models, thus facilitate the model reuse, model harmonization, and data integration. The objective of the study is to represent the HL7 detailed clinical models using OWL through leveraging the ISO 11179 standard, as case studies. In the process of the OWL-based transformation, we interpret the semantics between elements in DCM using the ISO 11179 standard and identify the relevant issues.

Background

W3C Semantic web recommendation The World Wide Web Consortium (W3C) is the main international standards organization for the World Wide Web. Its goal is to develop interoperable technologies and tools as well as specifications and

* Denotes equal contribution.

guidelines to lead the Web to its full potential. W3C recommendation has several maturity levels: Working Draft, Candidate Recommendation, Proposed Recommendation, and W3C Recommendation. When representing the HL7 DCM model, we considered using the following W3C recommendations: The Resource Description Framework (RDF) [11], RDF Schema [12], the Web Ontology Language (OWL) [7], OWL 2 [13], and Simple Knowledge Organization System (SKOS) [14]. In addition to these W3C recommendations, we also considered and included Dublin Core metadata element set (dc) [15], which are widely used to describe digital materials.

The ISO/IEC 11179 Standard The ISO/IEC 11179 standard, formally known as the ISO/IEC 11179 Metadata Registry (MDR) standard, is an international standard for representing metadata for an organization in a Metadata Registry [16]. The standard consists of six parts and the data element is foundational concept. The purpose of the standard is to maintain a semantically precise structure of data elements [17]. Each Data element in an ISO/IEC 11179 metadata registry: 1) should be registered according to the Registration guidelines; 2) will be uniquely identified within the register; 3) should be named according to Naming and Identification Principles; 4) should be defined by the Formulation of Data Definitions rules; 5) may be classified in a Classification Scheme.

HL7 Template and Archetype Architecture The HL7 Template Architecture (TA) [20] is a standard that specifies the syntax and semantics of the constituent components of clinical documents or messages for the purpose of exchange. The TA specification is richly expressive and flexible. As an example, document-level, section-level and entry-level templates can be used to constrain the generic Clinical Document Architecture specification. Archetype is a subset of templates, which syntactically and semantically structured aggregation of vocabulary or other data, which is the basic unit of clinical information. A formal language for expressing archetypes is known as Archetype Definition Language (ADL) described in [20]. The output of the Template process is a single formalism which could be represented in any suitable Knowledge Representation (KR) system including the Web Ontology Language (OWL). The mappings between ADL and OWL semantics have been discussed in the standard.

Table 1. The mappings between model constructs

HL7 DCM Construct	**Source**	**Example**	**OWL Construct**	**ISO 11179 Model Construct**
Variable	UML Class	Systolic blood pressure	A new OWL class	Data Element
Description for variable	Excel spreadsheet	The maximum pressure that is build in the aorta when the left ventricle contracts	dc:description	
Code for variable	UML attribute	271649006	annotation property or object property	Data Element Concept
Alternative code for variable	UML attribute		annotation property or object property	Data Element Concept
Datatype	attribute	PQ	OWL class	value_domain_datatype (attribute of Value Domain)
Example	attribute	140 mmHg	skos:example	Data Element Example
Vocabulary	Excel spreadsheet	SNOMED-CT	dc: source	
Method	UML attribute	Method valueset	Object property link to 0-1 value set	
Relationship between variables	UML association	Blood pressure	object property	
Valueset	UML attribute		A new OWL class. The allowed values are all subclasses of this class	Enumerated Value Domain
Value in valueset	enumeration		A set of OWL classes, sub class of the Valueset class	Permissible value
Description for value in a valueset	Excel spreadsheet		dc:description	
Code for value in valueset	Excel spreadsheet		Same as code for variable	Value meaning
Unit	attribute	in mmHg	muo:preferredUnit	value_domain_unit_of_ measurement (attribute of Value Domain)

OMG ODM - UML to OWL mapping Unified Modeling Language (UML) is a standardized general-purpose modeling language in the field of software engineering. The standard is managed, and was created by, the Object Management Group (OMG) [21]. The UML provides a graphical notation to express the design of object-oriented software system. The mappings of UML to RDF and OWL have been defined in the OMG Ontology Definition Meta-model (ODM) [22] and these enable the use of UML notation (and tools) for ontology modeling and facilitate generation of corresponding ontology descriptions in RDF, OWL respectively. This kind of transformation, however, usually is very general on the meta-data level, and therefore would not cover much on specific elements such as value sets, data ranges, data types, and etc.

Method

There are 10 HL7 DCM instances publicly available at HL7 wiki site [23]. Of them, we randomly selected 3 models for case studies: Blood Pressure, Body Height, and Body Temperature. For each model is described by an UML diagram for basic structure, an Excel Spreadsheet for representing the elements of model, and a Word documentation for describing the evidences of modeled domain.

The authors reviewed all available information about each model, identified the HL7 DCM constructs, and determined the mapping specifications between the model constructs and OWL constructs. Based on the specification, the model was represented in OWL manually using Protégé4 ontology editing environment [24]. The preliminary findings are discussed in the following sections.

Preliminary Findings and Discussion

Meta Model

We identified 14 HL7 DCM constructs from both UML diagrams of the models and their associated Excel spreadsheet as Column 1 in Table 1 shows These constructs form the basic meta-model which can be generalized to represent a target domain. Column 2 in Table 1 shows their corresponding sources for the constructs. We can see that some meta information such as description for variable, description for code, code for value, vocabulary are described in an Excel spreadsheet, rather than in an UML diagram.. Column 2 in Table 2 shows an example from the Blood Pressure model. Since the data type of Systolic blood pressure is defined as PQ, which means it should be an actual value. There is no value set associated to this variable. Therefore the corresponding cells are empty in the example.

After getting the HL7 DCM constructs identified, we found out that most of them can be directly mapped into the ISO 11179 data element model. The last column in Table 1 shows the mappings for the HL7DCM constructs with the ISO 11179 constructs. For instances, the DCM *Variable* is mapped as *Data Element* in the ISO 11179; the DCM *Code for variable* is mapped as *Data Element Concept*; the DCM *Valueset* is mapped as *Enumerated Value Domain*, and etc. In the case where the DCM does not have a formal or explicit definition for the constructs, we refer the 11179 model for a more accurate definition. For example, the DCM does not provide much definition for valuesets and its associated components. The ISO 11179, on the other hand, has a detailed UML diagram for *Enumerated Value Domain* (see Figure 1). We can refer the 11179 definition when representing the DCM in OWL.

OWL representation

We also map the DCM constructs using W3C constructs as the fourth column in Table 1 shows. Each HL7 DCM *Variable* or the ISO11179 *Data Element* is mapped to an OWL class. Its description is represented using *dc:description*. We proposed two different ways to represent codes to variables. One option is to use OWL annotation properties. We have created two OWL annotation properites: *prefRelatedCode* and *altRelatedCode* to represent the *Code for variable* and *Alternative code for variable* constructs in DCM. Using OWL annotation property provides a way for the model to keep track of the related code without further semantic assertions. OWL annotation properties, however, are not considered by reasoners. Another option is to use an OWL object property to link the reference codes and the variable itself. This way we can use the semantic definitions of the referenced concepts (i.e., a SCT concept) to define the variable class itself. Note that if the DCM chooses to define using post-coordinated SNOMED-CT concepts, this option would be a more suitable choice for capturing the semantic definition. For *Datatype*, each HL7 data type is defined as an OWL class. An object property called *HL7Datatype*, connects the class corresponding to the variable and its allowed data types. We adopted *skos:example* to represent possible examples for each variable, and *dc:source* to represent the source vocabulary. The DCM Method is represented by an OWL object property which links to zero to one value set. Relationships between variables are also represented using OWL object properties. We believe that the representations of valuesets and units are interesting enough to be discussed in separated sections.

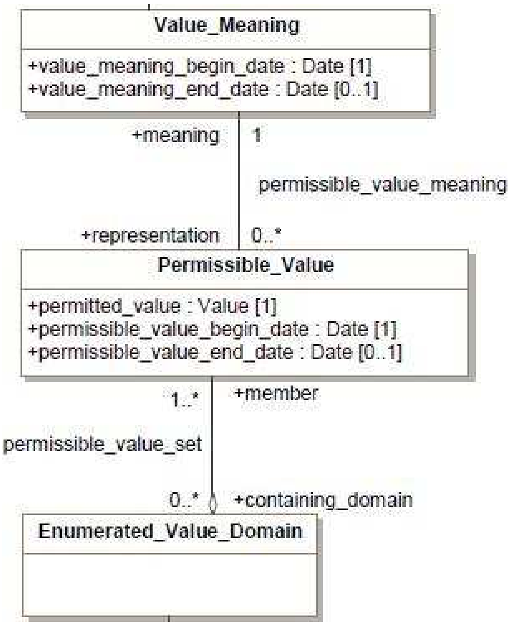
Value_Meaning
+value_meaning_begin_date : Date [1]
+value_meaning_end_date : Date [0..1]
+meaning
1
permissible_value_meaning
+representation
0..*
Permissible_Value
+permitted_value : Value [1]
+permissible_value_begin_date : Date [1]
+permissible_value_end_date : Date [0..1]
1..*
+member
permissible_value_set
0..*
+containing_domain
Enumerated_Value_Domain

Figure 1: The ISO 11179 Model for Value Set

Valueset

With the mappings into the ISO 11179 model, a valueset in DCM is interpreted as an enumerated value domain, each value in a valueset is interpreted as permissible value, and the code for a value in a valueset is interpreted as the meaning of the value. Figure 1 shows the ISO 11179 model for value sets. We adopted one of two standard representation approach proposed in the W3C [25], i.e. value partition approach in this study. In this approach, each valueset (i.e. enumerated value domain) is presented as an OWL class and the features of the class representing a continuous space that is partitioned by the values (i.e. permissible values) in the valueset. Note that HL7 common terminology service has a slightly different model for representing a valueset [26].

Each permissible value in the valueset is represented as individual OWL classes, and defined as sub-classes of the OWL class for the valueset itself. The OWL modeling decision for a permissible value and its value meaning (i.e. code for value) can be treated the same as that for a data element (i.e. variable) and its meaning - data element concept (i.e. code for variable), as described in the previous section.

Unit and Range

The DCM also defines the allowed units of measurement of variables when applicable. In order to ensure the semantic interoperability, we adopted the Measurements Unit Ontology (muo) [27] and the Unified Code for Units of Measure (ucum) [28] to describe the units of measurement for a specific DCM variable. For example, DCM specified that for the Body Height class, the allowed units could be either "cm" or "m". We can define two OWL classes called *bodyHeightinCM* and *bodyHeightinM* to represent the human body heights measured in "cm" and "m" respectively. We can further define the unit used and the data range for them using MUO properties and the OWL 2 data range assertions. For example, we can define the property *bodyHeightinCM* as follow:

```
:bodyHeightinCM rdf:type owl:DatatypeProperty ;
 rdfs:label        "Body Height in CM"@en ;
 muo:preferredUnit   ucum:cm ;
 rdfs:domain       BodyHeight ;
 rdfs:range        Minimum_and_maximum_of_body_height .
```

Note that the *preferredUnit* here is defined as an annotation property; therefore it is hard to make OWL reasoners to take it into consideration. Parsia and Smith have discussed the limitations and possible extensions of OWL to represent qualities in [29].

The Detailed Clinical Model can specify clinical content for use as Quality Measures. OWL2, a new version of OWL with extended features, is able to support the data range definition specified in DCM. OWL2 has a set of built-in numeric data ranges and provides the option for user to define data ranges using the basic built-in data ranges using expressive constructors [30]. For example, the DCM defined that the class "Body Height" has a dependency "minimum and maximum of body height", which defines the allowed data range of human body height. Using OWL2, we can define the data range as follow:

```
DatatypeDefinition(
 :Minimum_and_maximum_of_body_height
 DatatypeRestriction( xsd:decimal
  xsd:minInclusive "0.000"^^xsd:decimal
xsd:maxInclusive "1000.000"^^xsd:decimal  ) )
```

Conclusion

In this paper, we discussed our preliminary findings on the semantic harmonization of clinical study meta-data models such as the HL7 DCMs and the ISO 11179 model, under the framework of Semantic-Web technology. We first identified mappings between the DCM constructs with the constructs of the ISO 11179 model. In the case where the DCM does not have a formal or explicit definition for the constructs, we refer to the ISO 11179 model for a more accurate definition. We then used the Semantic-Web representation to represent the information presented in the DCMs. We consider that such as harmonization can provide computable semantics of the models, thus facilitate the model reuse, model harmonization and data integration.

Acknowledgements: This research is partially supported by the NCI caBIG Vocabulary Knowledge Center and the National Science Foundation under Grant #0937060 to the Computing Research Association for the CIFellows Project.

References

1. Warzel DB, Andonaydis C, McCurry B, Chilukuri R, Ishmukhamedov S, Covitz P. Common data element (CDE) management and deployment in clinical trials. AMIA Annu Symp Proc. 2003:1048. AMIA Annu Symp Proc. 2003
2. Komatsoulis GA, Warzel DB, Hartel FW, Shanbhag K, Chilukuri R, Fragoso G, Coronado S, Reeves DM, Hadfield JB, Ludet C, Covitz PA. caCORE version 3: Implementation of a model driven, service-oriented architecture for semantic interoperability. J Biomed Inform. 2008 Feb;41(1):106-23.
3. HL7 Detailed Clinical Model URL: http://wiki.hl7.org/index.php?title=Detailed_Clinical_Models.
4. Goossen WT. Using detailed clinical models to bridge the gap between clinicians and HIT. Stud Health Technol Inform. 2008;141:3-10.
5. Richesson RL, Krischer J.Data standards in clinical research: gaps, overlaps, challenges and future directions. J Am Med Inform Assoc. 2007 Nov-Dec;14(6):687-96. Epub 2007 Aug 21.
6. HL7 TermInfo Project URL: http://www.hl7.org/Special/committees/terminfo/index.cfm .
7. OWL overview: http://www.w3.org/TR/owl-features/.
8. Rector AL, Qamar R., Marley T. Binding Ontologies & Coding systems to Electronic Health Records and Messages. KR-MED 2006 "Biomedical Ontology in Action". November 8, 2006, Baltimore, Maryland, USA.
9. The W3C HCLS Clinical Observations Interoperability URL: : http://esw.w3.org/HCLS/ClinicalObservationsInteroperability.
10. Chen ES, Zhou L, Kashyap V, Schaeffer, M, Dykes PC, and Goldberg HS, Early Experiences in Evolving an Enterprise-Wide Information Model for Laboratory and Clinical Observations AMIA Annu Symp Proc. 2008:106-110
11. Resource Description Framework (RDF), http://www.w3.org/RDF/.
12. The RDF Schema vocabulary (RDFS), http://www.w3.org/2000/01/rdf-schema.
13. OWL 2, http://www.w3.org/TR/owl2-syntax/.
14. Simple Knowledge Organization System (SKOS), http://www.w3.org/TR/skos-reference/.
15. Dublin Core Metadata Element Set, http://dublincore.org/documents/dces/.
16. ISO/IEC 11179, Information Technology -- Metadata registries (MDR), http://metadata-standards.org/11179/.
17. ISO/IEC 11179 Standard article in WikiPedia: http://en.wikipedia.org/wiki/ISO/IEC_11179.
18. Beale T. Archetypes and the EHR. Stud Health Technol Inform 2003;96:238-44.
19. KILIC O., BICER V., DOGAC A., Mapping Archetypes to OWL, Technical Report TR-2005-3,http://www.srdc.metu.edu.tr/webpage/publications/2005/MappingArchetypestoOWLTechnical.pdf.
20. HL7 Template and Archetype Architecture, http://www.hl7.org/library/committees/template/HL7_Atlanta_10_20_04.doc.
21. Object Management Group, http://www.omg.org/.
22. Ontology Definition Metamodel (ODM), http://www.omg.org/spec/ODM/1.0/.
23. http://wiki.hl7.org/index.php?title=Detailed_Clinical_Models.
24. Protégé URL: http://protege.stanford.edu/.
25. Alan Rector, Representing Specified Values in OWL: "value partitions" and "value sets", http://www.w3.org/TR/swbp-specified-values/.
26. HL7 CTS Valueset Model: http://informatics.mayo.edu/LexGrid/downloads/CTS/specification/ctsspec/cts.htm#CTSValueSe.
27. Measurement Units Ontology, http://forge.morfeo-project.org/wiki_en/index.php/Units_of_measurement_ontology.
28. The Unified Code for Units of Measure, http://www.unitsofmeasure.org/.
29. Parsia B and Smith M, Quantities in OWL, OWLED 2008 OWL: Experiences and Directions.
30. OWL 2 Web Ontology Language Data Range Extension,, http://www.w3.org/TR/owl2-dr-linear/.