SEMANTIC INTEROPERABILITY BY ONTOLOGY-BASED REPRESENTATION OF CLINICAL INFORMATION

Martínez Costa C[1], Karlsson D[2], Schulz S[1,3]

Abstract

Semantic interoperability of clinical information requires establishing principles that guide what is represented by structural information models and terminologies in order to avoid their overlapping. Here we sketch a common framework based on ontologies and description logics for representing clinical information based on the hypothesis that the fully formalization of the overlapping area will improve semantic interoperability. By means of a simple use case we evaluate the framework by using some criteria for testing its information retrieval capability.

Keywords – EHR, Ontology, Terminology, DL, SNOMED CT, Structure, Boundary problem

1. Introduction

Despite the fact that free text continues being the most frequent way of representing medical data, there is a growing tendency to use medical terminologies and electronic health record (EHR) standards. Several terminologies have been used to encode important data of the medical record like problem lists or diagnoses. Coded data facilitate semantic interoperability, which means that data can be interpreted by systems that are different from those in which they were generated, and then used for different purposes, such as clinical research, decision support, health statistics, billing, etc. However, a set of codes is not enough to unambiguously represent the EHR. Codes appear in a specific context and some of them are interrelated. An example of this is a code like the ICD code E10.9 (*Insulin-dependent diabetes mellitus*) in the diagnosis or in the family history section of a patient record. The section in which the code is used provides the context and modifies the meaning. Structural models, such as EHR information models have to be used together with terminology models in order to fully represent the patient's EHR, seen as a collection of statements that encompass the observations about a particular patient, made by clinicians, each at a specific time and place for some purpose [1]. A clinical statement usually represents the information that answers to clauses characterized by interrogative pronouns like *what*, *how*, *when*, *where* and *who*: *who* stated it (e.g. clinician), *when* (date and time) and *where* (e.g. healthcare facility) represent the context or provenance of each statement and are essential for interpreting the information. *What* a

[1] Institute of Medical Informatics, Statistics and Documentation, Medical University Graz, Austria
[2] Department of Biomedical Engineering, Linköping University, Linköping, Sweden
[3] Institute of Medical Biometry and Medical Informatics University Medical Center, Freiburg, Germany

statement expresses can be represented by different combinations of information model structures with codes. For instance the context 'Family history' can be represented by combining information items with codes in the structural information model or by using the SNOMED CT 'Situation with explicit context' hierarchy, which offers single codes for frequent context / code combinations, like 430678008 for *Family History: Diabetes mellitus*. Due to this overlapping of information structure and domain terminology / ontology more than one representation for the same clinical information exists, which hampers semantic interoperability.

According to [2], interoperability can be seen from three points of views: (1) technical interoperability, (2) semantic interoperability, and (3) process interoperability. The first one focuses on moving data from a system A to a system B, without taking care of the data meaning. In contrast, semantic interoperability ensures that the two systems understand the data in the same way. Finally, process interoperability focuses on the semantic interoperability of clinical processes. The three kinds of interoperability are interdependent and necessary to obtain the full benefits of the EHR. In this work we will focus on semantic interoperability. In order to enhance semantic interoperability we need to detect the cases which represent the same semantic content in different ways. In order to compare such statements we represent them in a common formal framework. In Section 2 we present such a framework based on ontologies and description logics (DLs), which is capable of mediating between different combination of clinical information structures and clinical terminologies. Then, in Section 3 we provide a semantic interoperability use case and we use it to evaluate the framework in terms of data retrieval. Finally in Section 4 the conclusion and discussion are provided.

2. Methods

In [3], A. Rector et al. distinguished "model of meaning" from "model of use". Information based models as archetypes are defined as "models of use", since they describe how data are displayed or captured and are designed for specific use cases, while the "model of meaning" represents our understanding of the world, so that we can reason about it in general and about individual patients. Both are necessary since it is not viable that everybody agrees on the same way of capturing some information at the point of care (e.g. pick list vs. checklist). However, we are not interested in knowing the exact way some piece of information was captured but in knowing and communicating its meaning. Thus, in this work we are interested in the "model of meaning" and how it relates with the "model of use".

According to [4], (formal) ontologies are theories that attempt to give precise mathematical formulations of the properties and relations of certain entities in some domain. They can be seen as semantic building blocks for sharing information and knowledge across systems. Increasingly clinical terminologies, particularly SNOMED CT and the upcoming ICD-11, are based on formal ontologies. OWL, the Web Ontology Language [5] supported by the Protégé editor and several description logics reasoners has been established as a de facto standard in biomedical ontology research and practice The common framework we propose aims to detect semantically equivalent clinical information from iso-semantic models (structurally different but with same meaning). It consists of a common ontology described in OWL 2 DL, that includes top-level, domain and EHR information entity ontologies. Structural models will be represented by means of the information entity ontology while terminology models correspond to the domain ontology. Besides, we have used a top-level ontology in order to constrain and standardise how these two ontologies relate. As domain ontology we use parts of SNOMED CT [6], due to its increasing relevance as a comprehensive international EHR terminology with an increasing number of mappings to other classification systems. Then, the models of use (clinical models 1 and 2 in *Figure 1*) will be semantically enhanced with OWL DL expressions conforming to this common representation. As a result, each model will have a set of annotations that will be used to compute their equivalence by using DL reasoners such as HermiT[1], as depicted by *Figure 1*.

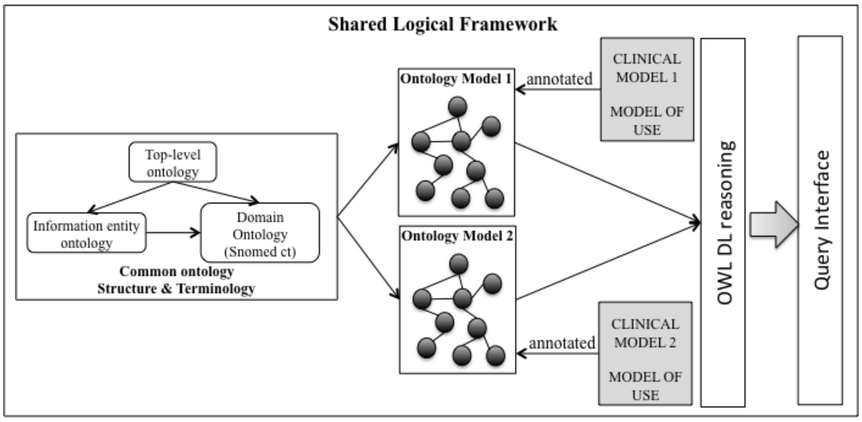

Figure 1: Shared logical framework schema

The OWL DL annotations will follow some predefined patterns and will be added to the empty clinical models during their creation. They will be later filled with the patient data given an appropriate software tool support. We are here mainly interested in representing information related to clinical situations like findings, disorders and diseases, for which we have created the relation isAboutSituation. It relates information entities from the information entity ontology with terminology content (ideally ontology-based such in SNOMED CT) that represents some patient situation (*InformationEntity* and isAboutSituation only *ClinicalSituation*). We approximate this linkage by using the universal quantifier 'only', due to the impossibility of OWL DL of directly targeting classes without quantifiers, which would correspond to second-order statements. The problem is that we do not want to logically assert existence and at the same time express that the existence of something is unclear, as with many clinical statements (e.g. the diagnostic information *Suspected encephalitis* does not allow to assert the existence of *Encephalitis* in that patient). According to [7] we propose to interpret the SNOMED CT finding concepts as clinical situations, e.g. as phases of a patient's life during which a certain condition of interest (structure, disposition, process) exists during all time. This interpretation has proved to be most robust and compatible with the existing hierarchies in terminologies like SNOMED CT and ICD. As an example, *Encephalitis situation* is any period of a patient's life during which he/she is the bearer of some encephalitis. The filler of the isAboutSituation relation may become arbitrarily complex according to the degree of postcoordination needed.

3. Results

Figure 2 depicts two snippets of heterogeneous and fictitious data entry forms used to gather data related to (suspected or confirmed) encephalitis. We do not make any assumption about the EHR standard used but we will use SNOMED CT as the common terminology. However, the clinicians may have selected different SNOMED codes for each form.

[1] Hermit OWL reasoner. http: //www.hermit-reasoner.com/.

We will evaluate the capability of the proposed framework to perform the information retrieval based on this use case and by using the following five criteria [2]:

- *Query expressivity*: the ability to support pre-coordination and/or post-coordination.
- *Subsumption testing*: the ability to test whether a patient with a specific condition can be assumed to have a more general condition that subsumes the specific one.
- *Context awareness*: the ability to take account of context data, recorded either in the record or in SNOMED expressions, when interpreting and evaluating results.
- *Concept equivalence*: the ability to retrieve equivalent information, when it has been recorded using different but equivalent terms and expressions.
- *Speed and performance*: the time and resources used in order to retrieve clinical data.

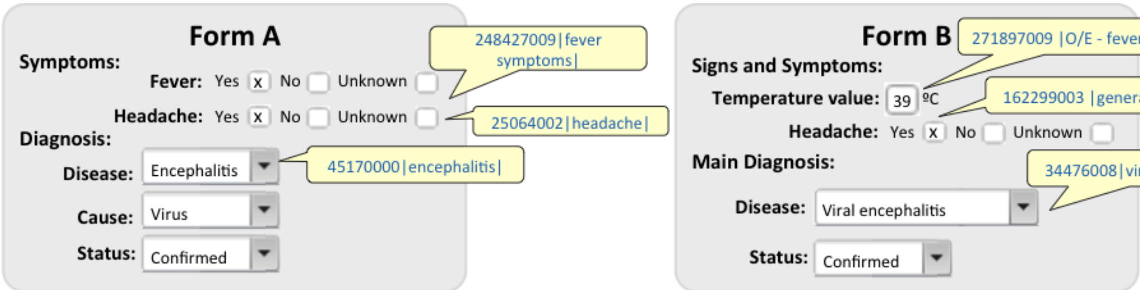

Figure 2: Two heterogeneous clinical forms for fictitious encephalitis use case and annotated with SNOMED CT

In order to retrieve both forms' data we will build a set of OWL DL annotations for each of the filled forms. These annotations are based on the common ontology proposed and will allow detecting the equivalences between the data from both applications with the help of an OWL DL reasoner.

3. 1. Query expressivity – Viral encephalitis confirmed diagnosis

In form A three fields named *Disease*, *Cause* and *Status represent* the patient's diagnosis. In form B only two fields, *Disease* and *Status* are used, and the first one is filled with the pre-coordinated concept *34476008* | *viral encephalitis* | which includes the cause of the disease. Each of the form fields will have associated an OWL DL annotation that corresponds to a defined class in the ontology (see *Tables 1 and 2*).

In order to compute the equivalence between the information represented by the forms, each one will correspond to the conjoint of its annotations. If we then run the DL reasoner, the conjunction of the annotations (A#2) and (A#3) are computed as equivalent to (B#2) and the query for all patients with viral encephalitis will subsume both of them.

Table 1: Form A. OWL DL annotations for *Viral encephalitis confirmed diagnosis*.

Form field	OWL DL annotation
Diagnosis (A#1)	*InformationItem* and **outcomeOf** some *DiagnosingProcess* and **isAboutSituation** only *Disease*
Disease: *Encephalitis* (A#2)	*Diagnosis* and **isAboutSituation** only *Encephalitis*
Cause: Virus (A#3)	*Diagnosis* and **isAboutSituation** only (*Disease* and **causedBy** *Virus*)
Status: Confirmed (A#4)	*Diagnosis* and **hasInformationObjectAttribute** some *Confirmed*

Table 2: Form B. OWL DL annotations for *Viral encephalitis confirmed diagnosis*.

Form field	OWL DL annotation
Main Diagnosis (B#1)	*InformationItem* and **outcomeOf** some *DiagnosingProcess* and **isAboutSituation** only *Disease*
Disease: *Viral Encephalitis* (B#2)	*Diagnosis* and **isAboutSituation** only *ViralEncephalitis*
Status: *Confirmed* (B#3)	*Diagnosis* and **hasInformationAttribute** some *Confirmed*

3. 2. Subsumption testing – Headache vs. Generalised headache

Although the fact that the patient has or not headache has been captured in the same way in forms A and B, two different SNOMED CT codes have been selected, *25064002 | headache |* and *162299003 | generalised headache |*. Since the second one is defined as a subclass of the first one in the terminology, both patients will be retrieved in case of asking for those that have headache. *Tables 3* and *4* show the OWL DL annotations added.

Table 3: Form A. OWL DL annotations for *Headache*

Form field	OWL DL annotation
Headache: Yes (A#1)	*InformationItem* and **isAboutSituation** only *Headache*

Table 4: Form B. OWL DL annotations for *GeneralisedHeadache*

Form field	OWL DL annotation
Headache: Yes (B#1)	*InformationItem* and **isAboutSituation** only *GeneralisedHeadache*

3. 3. Context awareness – Fever

In form A the patient fever is recorded by using a check list (*Yes, No*). In form B is recorded by using a field named *Temperature value* that allows to specify the patient temperature in ºC. Clinicians have used different codes for the fever, the concept *248427009 | fever symptoms |* in form A and *271897009 | O/E fever |* in form B. In SNOMED CT these concepts are not related by taxonomic subsumption; therefore a query for patient with fever would not retrieve both of them. If we look up into the concept definition in SNOMED CT their meaning is slightly different. In form A, the fact that the patient has fever is the result of asking the patient, while in form B it is the result of some physical examination (probably measuring fever by a health professional using a thermometer). *Tables 5 and 6* show the annotations added to each form.

Table 5: Form A. OWL DL annotations for *Fever*.

Form field	OWL DL annotation
Fever: Yes (A#1)	*InformationItem* and **isAboutQuality** only *Fever* and **outcomeOf** some *HistoryTakingProces*s

As it can be derived from the annotations added, in both cases we are referring to the fact that the patient has fever. However, in form A the fever is a symptom told by the patient to the clinician while in form B it is inferred from the temperature value obtained after some physical exam, since Fever is defined as:

Fever **equivalentTo** *Temperature* and (**inheresIn** some *CorePartBody*)
and (**hasValue** some integer[> 37])

With this reinterpretation, the query for obtaining all patients with mention of fever will get both of them. The above definition of *Fever* is essential for inferring that for the patient from system B also a fever situation is mentioned in the record. This is not provided by SNOMED CT and therefore additional efforts will be needed in order to encode it. Besides, fever might be differently represented by other systems (i.e. core body temperature > 37.2). Therefore, a desideratum is that terminological standards like SNOMED CT should provide formal or informal definitions of what is exactly meant by a concept.

Table 6: Form B. OWL DL annotations for *Temperature value 39*.

Form field	OWL DL annotation
Temperature value: 39 (B#1)	*InformationItem* and **isAboutQuality** some *Temperature* *InformationItem* and **isAboutQuality** only (*Temperature* and **hasValue** value 39 and **inheresIn** some *CorePartBody*) and **outcomeOf** some *PhysicalExaminationProcess*

3. 4. Concept equivalence

This criterion is described as the ability to retrieve equivalent information, when it has been recorded using different but equivalent terms and expressions and as we have shown above it is accomplished by all the examples provided.

3. 5. Speed and performance

Performance issues are crucial when automated reasoning is used. Description logics reasoning can only be built into routine systems if it is assured that the reasoning steps happen in polynomial time. As a consequence, the expressivity of logics must be kept simple as it is the case with SNOMED CT, which subscribes to the OWL 2 EL profile, a rather inexpressive variant of OWL DL, which lacks negation, disjunction and universal restrictions. Assuming that in SNOMED CT clinical findings are interpreted as clinical situations, as introduced above negation and disjunction will be needed in some cases, which means to move it to the OWL DL representation. In order to keep performance within an acceptable time frame the extraction of terminology subsets for specific use cases as well as other targeted ontology simplification measures might be essential. This constitutes a large area of further investigations and simulations. It is not part of the work presented here, which aims at framing the problem from an ontological point of view as an initial step.

4. Discussion and conclusion

There are many projects and initiatives that pursue the improvement of semantic interoperability of clinical information. Examples of these are CIMI[1], LRA[2], SIAMM[3], CEM[4] or DCMs[5]. The work presented here has been developed in the context of the SemanticHealthNet EU project. It differs from the above proposals in the sense that it tries to be less prescriptive and more descriptive. It

[1] http://informatics.mayo.edu/CIMI/

[2] http://www.connectingforhealth.nhs.uk/systemsandservices/data/lra

[3] G. Freriks ERS B.V. in collaboration with the EN13606 Association. To be published and sunmitted to CEN/ISO

[4] SHARPn project: http://informatics.mayo.edu/sharp/

[5] http://www.detailedclinicalmodels.nl/home-en

assumes that present as well as future systems represent clinical information using different EHR models (proprietary or based on some standard), and that the cost of implementing a map to a "canonical" model that supports semantic interoperability is too high. Therefore, we aim at providing a semantic layer that will sit between these information systems and a sophisticated query system. This layer will allow providing a consistent semantic representation of clinical data and detecting semantic equivalences among them. It consists of the common formal framework described. Then, the "models of use" will be semantically enhanced with expressions conforming to the common representation proposed. These expressions are represented in OWL DL. Thus, a sophisticated query system will use the results from the semantic layer in order to build queries for retrieving data from heterogeneous medical repositories.

However, the costs of providing this semantic layer do not have to be underestimated. The semantically enhancement of the "models of use" is not a trivial task. At this moment it has to be manually done, however, by using the appropriate patterns and with the support of informatics tools we think that the approach could be feasible. Evidence of this should be one of the outcomes of the SemanticHealthNet project.

5. Acknowledgments

This work has been funded by the SemanticHealthNet Network of Excellence within the EU 7th Framework Program, Call:FP7-ICT- 2011-7, agreement 288408. http://www.semantichealthnet.eu/

6. References

[1] Rector A, Nowlan W, Kay S. Foundations for an electronic medical record. Methods Inf Med.1991;30:179-86.

[2] Benson T. Principles of Health Interoperability HL7 and SNOMED - Health Information Technology Standards, 2012, p.101-119.

[3] Rector A, Qamar R, Marley T. Binding Ontologies & Coding systems to Electronic Health Records and Messages, Applied Ontology. 2009;4:51-69.

[4] Hofweber T. Logic and Ontology, Stanford Encyclopaedia of Philosophy, 2004. Available from: http: //plato.stanford.edu/entries/logic-ontology. Last accessed: April 2013.

[5] W3C OWL working group. OWL 2 Web Ontology Language, Document Overview. W3C Recommendation 11 December 2012 http: //www.w3.org/TR/owl2-overview/. Last accessed: April 2013.

[6] Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT), 2008. Available from: http: //www.ihtsdo.org/snomed-ct. Last accessed: April 2013.

[7] Schulz S, Rector A, Rodrigues J-M, & Spackman K. (Competing Interpretations of Disorder Codes in SNOMED CT and ICD. In: Proc. of AMIA 2012 Annual Symposium, 2012, Chicago, U.S.

Corresponding Author
Catalina Martínez Costa
Institut für Medizinische Informatik, Statistik und Dokumentation, Medizinische Universität Graz
Auenbruggerplatz 2/V , 8036 Graz
Email: catalina.martinez@medunigraz.at