Ontology Content Patterns as Bridge for the Semantic Representation of Clinical Information

Catalina MARTÍNEZ-COSTA[a,1], Stefan SCHULZ[a,b]
[a] *Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Austria*
[b] *Institute of Medical Biometry and Medical Informatics, Freiburg University Medical Center, Germany*

Abstract. Semantic interoperability of the Electronic Health Record (EHR) requires a rigorous and precise modelling of clinical information. Our objective is to facilitate the representation of clinical facts based on formal principles. We here explore the potential of ontology content patterns, which are grounded on a formal and semantically rich ontology model and can be specialised and composed. We describe and apply two content patterns for the representation of data on tobacco use, rendered according two heterogeneous models, represented in openEHR and in HL7 CDA. Finally, we provide some query exemplars to show a data interoperability use case.

Keywords. Electronic Health Records, Semantics, SNOMED CT, Knowledge Representation

1. Introduction

The notion of clinical model patterns has become popular in activities targeting the semantic interoperability of electronic health records (EHRs) [1][2]. They are design patterns that address recurrent modelling issues and are related to information models, which they constrain by following certain rules, and for which they create content definitions for use cases like 'acute care summary' or 'radiology report'.

Design patterns should keep separate the *model of use* from the *model of meaning* [3]. Different combinations of information model structures will often produce different models of use with the same meaning, so-called iso-semantic models. Whereas information models ideally constitute the (epistemic) model of use, the domain terminologies constitute the (ontological) model of meaning. These models should complement each other, but in practice there are overlaps, which complicate the identification of iso-semantic content.

In this work, we introduce ontology content patterns for representing clinical information based on a formal reference model underpinned by ontological principles, which allows providing clinical information with precise semantics, and thus paves the way to compute the equivalence between syntactically different but semantically same

[1] Corresponding Author: catalina.martinez@medunigraz.at

expressions. As much as it would be desirable that such patterns provide rigid principles to encode clinical information, we have to admit that a single way of encoding a given piece of information cannot be enforced. The EU SemanticHealthNet (SHN) network [4] addresses this problem by proposing a semantic infrastructure based on an ontological framework [5], together with a set of ontology content patterns [6] that use this framework as a reference. The framework consists of three kinds of ontologies: (i) top-level; (ii) information entity and (iii) medical domain, expressed in OWL 2 DL [7]. How this framework interacts with content patterns will be explained in the following.

We provide a subset of top-level ontology content patterns represented as subject-predicate-object (SPO) triples. By means of their specialization they capture the semantic representation of typical clinical information. Our interoperability use case focuses on two heterogeneous clinical models rendered in openEHR and in HL7 CDA.

Finally we show some query exemplars to briefly describe some of the benefits of the representation of the clinical information according to the framework proposed.

2. Methods

2.1. Ontological Framework

Ontology content patterns are based on a set of related ontologies which conform the SHN ontological framework. In our case it consists of three ontologies:

- A top-domain ontology, BioTopLite [10] (prefix btl:) providing a set of canonical top-level classes and relationships, like *btl:Condition*, *btl:InformationObject*, *btl:Quality*, *btl:Process* or **btl:hasPart**, **btl:bearerOf**, respectively.
- A domain ontology, SNOMED CT [11] (prefix sct:), a huge clinical terminology partially built on formal-ontological principles. Selected SNOMED CT content will be placed under top-level classes provided by BioTopLite.
- An EHR information entity ontology (prefix shn:) for representing pieces of information like diagnostic statements, plans, orders, etc. They are outcomes of clinical actions like observations, investigations, or evaluations. All classes of this ontology are represented as subclasses of the top-level class *btl:InformationObject*.

Information entities will refer to (types of) clinical entities by means of the relation **btl:represents** which can be further specialized by **shn:isAboutSituation** and **shn:isAboutQuality** for referring to a patient clinical situation [12] or a quality indirectly or directly observed of some material object or process.

2.2. Content Patterns

Ontology content patterns provide a particular view on ontology, tailored to the needs of particular use cases [13]. They can be organized in hierarchies, in which specializations follow a similar paradigm to the object-oriented design, and in which their composition permits to cover larger modelling use cases [14]. We propose the use of ontology content patterns as a "proxy" which allows representing clinical information according to the ontology-based representation previously described and prevents users from a deep knowledge of ontology and description logics syntax.

Our assumption is that a broad range of clinical models can be represented by the specialisation and composition of a limited set of ontology content patterns. In [15], we demonstrated the creation and application of such patterns for representing information on *heart failure* in a bottom-up approach. We found out that they could be described by means of specialisation and composition based on a set of higher-level patterns (top-level patterns). Here, we describe two top-level patterns and demonstrate their use for representing clinical information from two clinical models on *tobacco use*. The patterns are encoded as SPO triples, enhanced by a cardinality attribute. Note that the predicates are defined at the level of the pattern and are not taken from the source ontologies. They constitute direct links between classes, whereas OWL DL object properties only connect individuals. Top-level patterns can be specialized and composed by following certain cardinality and value restrictions. On the one hand, cardinality constraints place a constraint on the number of instances in which some predicate is used with different values. Note that at this level the instances are object classes, not domain individuals). Value range constraints limit the possible values for some predicate, allowing another pattern as object part of a triple.

The first top-level pattern we will describe (see Table 1) can be used to represent some piece of information about a particular clinical situation of the patient. Clinical situations, as described in [12], correspond to SNOMED CT findings. The first table row provides the clinical situation in focus. The second row represents the process performed to acquire the information. (e.g. diagnostic, physical examination, history taking, etc.) Finally, the third row specifies any information aspect related with the clinical situation in focus (e.g. severity, certainty, etc.)

Table 1 Information about Clinical Situation Content Top-Level Pattern

#N	Subject	Predicate	Cardinality	Object
S1	*shn:InformationItem*	'describes situation'	1..*	*shn:ClinicalSituation*
S2	*shn:InformationItem*	'results from process'	1..*	*btl:Process*
S3	*shn:InformationItem*	'has attribute'	0..*	*shn:InformationAttribute*

The second top-level pattern (Table 2) is the observation result pattern which describes the result of an observation or assessment about some quality of a given clinical situation. The first two rows describe the quality observed / assessed and the clinical situation, respectively. The third, fourth and fifth rows describe the result of the observation / assessment; the sixth row the scale in which the observed value is based. Finally, the last row represents the process performed to acquire the information.

Table 2 Observation result about process quality Content Top-Level Pattern

#N	Subject	Predicate	Cardinality	Object
O1	*shn:ObservationResult*	'describes quality'	1..1	*btl:Quality*
O2	*btl:Quality*	'is quality of'	1..*	*shn:ClinicalSituation*
O3	*shn:ObservationResult*	'has observed value'	1..1	*btl:ValueRegion*
O4	*btl:ValueRegion*	'has value'	0..1	*xml:datatype*
O5	*btl:ValueRegion*	'has units'	0..1	*shn:MeasurementUnits*
O6	*btl:ValueRegion*	'has scale'	0..1	*shn:Scale*
O7	*shn:ObservationResult*	'results from process'	1..*	*btl:Process*

2.2.1. OWL DL representation

The representation of the above top-level patterns into OWL 2 DL allows the precise formalization of the ontological framework proposed and the use of DL reasoning. DL reasoning is useful for the achievement of two important goals: On the one hand, it can

be used for detecting equivalent clinical information from iso-semantic models [16]. This includes the ability to compare different distributions of content between information models and ontologies/terminologies, in order to test whether they are semantically equivalent. For instance, there are two possible representations to encode a breast cancer diagnosis when using SNOMED CT: (1) using one diagnosis information model element and the concept *Breast cancer* or (2) using two information model elements for representing the disease diagnosed *Cancer* and the disease location *Breast structure*. An appropriate representation, supported by a DL reasoner should discover that both representations are semantically equivalent.

In our use case, DL reasoning can be used to provide an advanced exploitation of clinical information by means of semantic query possibilities such as retrieving patients who use tobacco, independently of the form of the tobacco (e.g. cigar, pipe, etc.) and of the type of consumption (e.g. snuff or smoking).

Table 3 depicts the translation of the patterns into OWL DL, according to the proposed ontological framework. By following the triple-based pattern representation shown in Tables 1 and 2, the subject (SUB) and object (OBJ) correspond to ontology classes and the predicate to an OWL DL expression. These DL expressions use one or more object properties from our ontologies, together with different quantifier, as a result of the underlying ontological model. In case the latter is modified, the change can be performed at this place, whereas the pattern representation remains the same.

Table 3. OWL DL representation of the top-level patterns

Predicate	**OWL DL expression**	
'describes situation'	SUBJ	subClassOf **shn:isAboutSituation** only OBJ
'describes quality'	SUBJ	subClassOf **shn:isAboutQuality** only OBJ
'results from process'	SUBJ	subClassOf **btl:isOutcomeOf** some OBJ
'has attribute'	SUBJ	subClassOf **btl:hasInformationAttribute** some OBJ
'is quality of'	SUBJ	subClassOf **btl:inheresIn** some OBJ
'has observed value'	SUBJ	subClassOf *btl:Quality* and **btl:projectsOnto** some OBJ
'has value'	SUBJ	subClassOf **btl:isRepresentedBy** only (**shn:hasInformationAttribute** some OBJ)
'has units'	SUBJ	subClassOf **btl:isRepresentedBy** only (**shn:hasValue** some OBJ)
'has scale'	SUBJ	subClassOf **btl:isRepresentedBy** only (**shn:hasInformationAttribute** some OBJ)

2.2.2. OpenEHR and HL7 CDA tobacco use models

We apply these patterns to an excerpt of an HL7 CDA and an openEHR model, which describe information about the patient's tobacco consumption. Each one had been designed by different requirements and for different contexts.

The openEHR model is part of the heart failure summary, developed by SHN, using the openEHR representation available in the Clinical Knowledge Manager (CKM) [17]. It collects detailed information about tobacco consumption, obtained from different sources, targeted to investigate the tobacco use in heart failure patients.

The HL7 CDA model follows one of the templates defined as part of the Consolidated CDA (C-CDA) solution [18] which provides a library of reusable CDA templates. The template comprises the data elements and vocabulary requirements needed for meeting the EHR Certification Criteria in support of the U.S. *Meaningful Use Stage 2* [19] and might be extended depending on additional information requirements. Thus, this CDA model is very generic and only records a patient's

smoking status within the social history section of the patient record. Table 4 shows an excerpt of some data elements and terminology value requirements of either model.

Table 4. Data elements and values (SNOMED CT) of an excerpt of openEHR and HL7 tobacco models

openEHR			**HL7 CDA**		
Data Element	**Value**		**Data Element**	**Value**	
Smoking status	77176002 8392000 8517006 160616005	*Smoker* *Non-smoker* *Ex-smoker* *Trying to give up smoking*	Smoking status	449868002 428041000124106 8517006 266919005 428071000124103	*Current every day smoker* *Current some day smoker* *Former smoker* *Never smoker* *Heavy Tobacco Smoker, etc.*
Form	<<39953003	*Tobacco*			
Typical smoked amount	259032004	*Quantity and units per day*			

The openEHR model records: the current tobacco smoking activity (e.g. *Current tobacco smoker*); the form of the tobacco (e.g. cigarette, in the above table "<<" means all subclasses) and the typical tobacco amount per day (e.g. 10 cigarettes). The HL7 CDA model provides only a data element for recording the tobacco smoking status. The status value is constrained to a set of SNOMED CT codes to meet the certification criteria in support of *Meaningful Use Stage 2* (e.g. *Current every day smoker*).

3. Results

In order to get the semantic representation of some fictitious clinical data rendered according the openEHR and HL7 CDA models, we have to (i) specialize/compose the top-level patterns described in section 2 and (ii) establish the correspondences between the model data element / value pairs and the pattern triples. As clinical data examples we will represent the following pairs (cf. Table 4): OpenEHR: Smoking status*/Smoker* (*77176002*); Form*/Cigarette smoking tobacco* (*66562002*); and typical smoked amount/*10 per day*; HL7 CDA: Smoking status*/Heavy cigarette smoker* (*230063004*).

Some SNOMED CT terms are misleading. E.g., *Smoker* does not refer to a person but to a smoking situation since it is placed in the clinical finding hierarchy. Thus, the use of the same term with different meanings by the EHR systems will hamper semantic interoperability. The knowledge model they conform to can be used to determine the real meaning of the term. However this model might be faulty or incomplete as it happens with the terms *Cigarette smoking tobacco* and *Cigarette tobacco smoker*, which refer to a substance and finding, respectively, without providing any relationship between both. Therefore, there will be no interoperability if systems use both of them arbitrarily. Table 5 shows the code and full specified name (FSN) of the SNOMED CT terms we use in the upcoming examples and our suggested re-naming based on their parent concepts.

Table 5. Meaning and renaming of the SNOMED CT concepts (ID and fully specified name)

SNOMED CT CODE & FSN	**Renaming suggestion**
77176002 *Smoker* (finding)	*tobacco smoking situation*
66562002 *Cigarette smoking tobacco* (substance)	*cigarette tobacco smoke substance*
65568007 *Cigarette smoker* (finding)	*cigarette tobacco smoking situation*
230063004 *Heavy cigarette smoker* (finding)	*heavy cigarette tobacco smoking situation*

Next, we show the top-level patterns specialisation required to represent the clinical data examples and provide the correspondences between the patterns and the openEHR and HL7 CDA models. Finally we describe some query exemplars on the data.

3.1. Semantic representation of the openEHR clinical data

Table 6 depicts the specialisation of the top-level ontology content patterns from Section 2 in order to represent the openEHR conforming clinical data. The left and right columns show the correspondences between the model data elements / value pairs and the pattern triples. The smoking status and the form are both mapped to the *Information about clinical situation* pattern, since the smoking status refers to a patient smoking situation and the form is part of the situation class definition, refining it. The typical amount smoked is mapped to the *Observation result pattern* since it is an assessment result. In the same table, the triples obtained are provided. Triples with minimum cardinality one are mapped to the model (eg. *shn:InformationItem* 'describes situation' *shn:ClinicalSituation).* Value constraints have been applied constraining the object part of the triple (e.g. *shn:InformationItem* 'describes situation' *shn:ClinicalSituation*) to the specific clinical situation (*sct:TobaccoSmokingSituation*).

Table 6. OpenEHR: "Smoker, cigarette smoker, 10 cigarettes per day"; Correspondences and Pattern triples

Data Element / Value	Triple representation	#N
Smoking Status / *smoker* (finding)	*shn:InformationItem* 'describes situation' *sct:TobaccoSmokingSituation* *shn:InformationItem* 'results from process' *sct:HistoryTaking*	#S1 #S2
Form / *cigarette smoker* (finding)	*shn:InformationItem* 'describes situation' *sct:CigaretteTobaccoSmoking Situationshn:InformationItem* 'results from process' *sct:HistoryTaking*	#S1 #S2
Typical smoked amount / *10 cigarettes / day*	*shn:ObservationResult* 'describes quality' *shn:MassIntake* *shn:MassIntake* 'is quality of' *sct:CigaretteTobaccoSmokingSituation* *shn:ObservationResult* 'has observed value' *btl:ValueRegion* *btl:ValueRegion* 'has value' *10* *btl:ValueRegion* 'has units' *sct:PerDay*	#O1 #O2 #O3 #O4 #O5

3.2. Semantic representation of the HL7 CDA clinical data

Table 7 depicts the result of specialising the top-level content patterns and the correspondences with regards to the HL7 CDA data. The smoking status, as in the openEHR case, is mapped to the information about clinical situation pattern.

Table 7. HL7 CDA "Heavy cigarette tobacco smoker (>=10)"; Correspondences and Pattern triples

Data Element / Value	Triple representation	#N
Smoking Status / *Heavy Cigarette Tobacco Smoker*	*shn:InformationItem* describes situation *sct:HeavyCigaretteSmokingSituation* *shn:InformationItem* results from process *sct:Evaluation*	#S1 #S2

The HL7 CDA model defines heavy smoker as at least 10 cigarettes / day. However, the definition is particular to this HL7 implementation and might vary across institutions or depend on research study purposes.

3.3. Querying the semantic representation of the openEHR and HL7 CDA clinical data

Table 7 depicts DL query exemplars based on the OWL DL representation of the openEHR and HL7 CDA data. The triple-based representation is transformed into OWL according to Table 3. We have formulated the following queries, asking at different information granularity level: (Q1) information about tobacco smokers; (Q2) information about heavy smokers; (Q3) information about cigarette smokers and heavy smokers; (Q4) information about patients which smoke more than 15 cigarettes / day.

Table 7. DL Query examples

#Q1	*shn:InformationItem* and **btl:isOutcomeOf** some *sct:HistoryTaking* and **shn:isAboutSituation** *only sct:TobaccoSmokingSituation*
#Q2	*shn:InformationItem* and **btl:isOutcomeOf** some *shn:Evaluation* and **shn:isAboutSituation** *only sct:HeavyTobaccoSmokingSituation*
#Q3	*shn:InformationItem* and **btl:isOutcomeOf** some *shn:Evaluation* and **shn:isAboutSituation** *only sct:HeavyTobaccoSmokingSituation* and **shn:isAboutSituation** *only sct:CigaretteTobaccoSmokingSituation*
#Q4	*shn:ObservationResult* and **shn:isAboutQuality** only (*shn:MassIntake* and **btl:inheresIn** some *sct:CigaretteTobaccoSmokingSituation* and **btl:projectsOnto** some (*btl:ValueRegion* and **btl:isRepresentedBy** only (**shn:hasInformationAttribute** some *sct:PerDay* **shn:hasValue** some int[>15])))

The four queries use DL reasoning. Q1 ask for tobacco smokers. It will retrieve both openEHR and HL7 CDA like data since a *Heavy tobacco smoking situation* is a subclass of a *Tobacco smoking situation*. Q2 ask for heavy smoker without specifying the form. It retrieves both data instances, since *Heavy cigarette smoker* is a subclass of *Heavy smoker* and we have defined that a *Heavy cigarette smoker* means at least 10 cigarettes / day and this is the typical smoked amount provided by the OpenEHR data. Q3 specifies the query asking by those who are heavy smokers and smoke using cigarettes, which is the same as asking for *Heavy tobacco cigarette smoking situation*. Finally, Q4 ask by those whose typically smoke more than 15 cigarettes / day, and do not retrieve anything, since they smoke 10 / day.

4. Discussion and Conclusion

From the above we can state (i) that it is not possible to impose a single model representation across diverse clinical communities (e.g. public health vs. primary care vs. specialised care) and clinical practices, and (ii) that the requirements will dictate the level of information detail needed. Then, by considering these clinical limits, the immediate question is which degree of semantic interoperability we can offer, or up to which degree we can make the above models semantically interoperable.

SemanticHealthNet (SHN), in contrast to other proposals does not intend to provide a new EHR standard. Instead it provides an intermediate semantic layer able to deal with the unavoidable heterogeneity which arises when clinical information is represented across or within the same medical domain. SHN's semantic infrastructure is based on an ontological framework and a set of ontology content patterns that use

this framework as a reference. It proposes to use ontology content patterns to assist in information modelling, preventing the user from fully understanding the underlying, complex, ontological expressions. Content patterns can be specialised and composed to cover the needs of different use cases. They do it by following a formal framework and a set of constraints which keep them semantically interoperable. One of the main research questions which have still to be investigated is whether there are a finite number of top-level patterns from which the others will specialize. We only can argue that representation of the information on heart failure [15] provided a high degree of information heterogeneity and that a reduced number of top-level patterns were derived from that. Besides, the technological uptake of this approach will require a series of challenges (human, computational) to be met. We think that human challenges such as the ontology-based representation of present clinical information could be alleviated by using semantic artefacts such as ontology content patterns, which might be implemented by specific tools. However, computational challenges in most cases require the evolution of present tools and resources, which might lead to agree on compromises between performance and functionality.

References

[1] S. Heard, T. Beale, G. Freriks, A. R. Mori, O. Pishec. Templates and Archetypes: How do we Know What We are Talking About, Version 1.2, 2003

[2] CIMI Patterns: http://informatics.mayo.edu/CIMI/index.php/ (Last accessed: Jan 2014).

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

[4] SemanticHealthNet Network of Excellence. http://www.semantichealthnet.eu/ (Last accessed Jan. 2014)

[5] S. Schulz, C. Martinez-Costa. How Ontologies Can Improve Semantic Interoperability in Health Care. Lecture Notes in Computer Science Volume 8268, 2013, 1-10.

[6] A. Gangemi, V. Presutti, Content Ontology Design Patterns As Practical Building Blocks for Web Ontologies. In Proc. of the 27th International Conference on Conceptual Modeling, 2008, 128-141

[7] 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 Jan. 2014)

[8] OpenEHR. An open domain-driven platform for developing flexible e-health systems. http://www.openehr.org (Last accessed Jan. 2014)

[9] R. H. Dolin, L. Alschuler, S. Boyer, C. Beebe, F. M. Behlen, P.V. Biron, A.S. Shvo. The HL7 Clinical Document Architecture, release 2. JAMIA. 2006;13:30-39

[10] BioTopLite: http://purl.org/biotop/biotoplite.owl. (Last accessed Jan. 2014)

[11] Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT), 2008. http://www.ihtsdo.org/snomed-ct (Last accessed Jan. 2014)

[12] S. Schulz, A. Rector, J. Rodrigues, C. Chute, B. Üstün, K. Spackman, Ontology-based convergence of medical terminologies. SNOMED CT and ICD-11. Proc. of eHealth2012. Vienna, Austria: OCG, 2012.

[13] E. Blomqvist, E. Daga, A. Gangemi, V. Presutti, Modelling and using ontology design patterns. [http://www.neon-project.org/web-content/media/book-chapters/Chapter-12.pdf]

[14] A. Gangemi, Ontology Design Patterns for Semantic Web Content. In Proceedings of the Fourth International Semantic Web Conference, 2005, pp. 262-276

[15] SemanticHealthNet deliverable 4.2: Ontology/Information models covering the HF use case, 2013.

[16] C. Martínez Costa, D. Boscá. M.C. Legaz-García, C. Tao, J.T. Fernández-Breis, S. Schulz, C.G. Chute, Isosemantic rendering of clinical information using formal ontologies and RDF. In Proc. MEDINFO 2013. Stud Health Technol Inform. 2013;192:1085

[17] Clinical Knowledge Manager. http://www.openehr.org/ckm/ (Last accessed Jan. 2014)

[18] HL7 IG for CDAR2: IHE Health Story Consolidation, R1", Consolidated CDA, C-CDA: http://www.hl7.org/implement/standards/ (Last accessed Jan. 2014)

[19] US Meaningful Use Stage 2: http://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/Downloads/Stage2_Guide_EPs_9_23_13.pdf (Last accessed Jan. 2014)