Improving EHR Semantic Interoperability Future Vision and Challenges

Catalina MARTÍNEZ-COSTA[a,1] Dipak KALRA[b], Stefan SCHULZ[a]
[a] *IMI,Medical University of Graz, Austria*
[b] *CHIME, University College London, UK*

Abstract. We propose a semantic-driven architecture to improve EHR semantic interoperability as result of the work carried out within the SemanticHealthNet project. The architecture spans from structured data standardized or not, to their use by heterogeneous application systems. Each of the architecture layers is described as well as the challenges of implementing them in practice. These challenges are mostly described from the technical side and not from the human one, although the latter are also important.

Keywords. EHR, semantic interoperability, SNOMED CT, ontology

Introduction

Shared clinical terminologies and ontologies as a means to enable the faithful exchange of the meaning of information had been addressed in the EU SemanticHealth interoperability roadmap [1], but their interplay and their interfaces still constitutes a major desideratum. We here address this issue by proposing a semantic-driven architecture. It is organized in five layers and spans from heterogeneous data repositories to homogeneous and semantically explicit representations. The work is part of SemanticHealthNet (SHN) [2] , which has focused, as a use case, on the interoperable representation of heart failure information to support integrated care. The architecture presented here has been drawn from, and validated in this clinical domain.

In the following we explain some of the challenges in each of the layers of this architecture. Some of them are technical and derived from the use of semantic technologies with any data management system. Others are specific to the medical domain and the use of EHR standards and medical terminologies.

1. Vision of an EHR semantic-driven architecture

Structured heterogeneous data (Layer 1): This layer comprises structured clinical data, which may be physically stored within an EHR repository and accessed via an interface conforming to some standard like HL7 CDA, openEHR, EN ISO 13606, or to a proprietary database schema. Advanced EHR representations structure data by clinical models (e.g. archetypes), which provide a set of data elements and value restrictions. More than one expression [*Data element*: *Value restriction*] can be used for represent-

[1] catalina.martinez@medunigraz.at

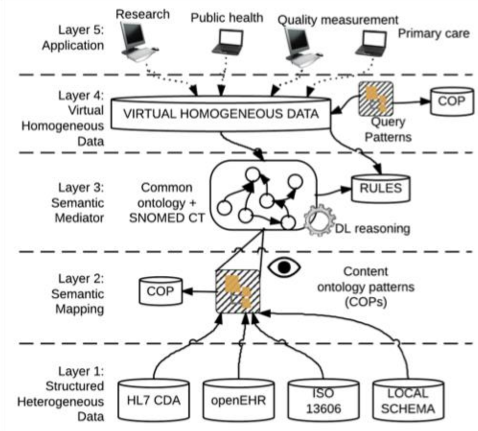
Layer 5: Application
Research
Public health
Quality measurement
Primary care
Layer 4: Virtual Homogeneous Data
VIRTUAL HOMOGENEOUS DATA
COP
Query Patterns
Layer 3: Semantic Mediator
Common ontology + SNOMED CT
RULES
DL reasoning
Layer 2: Semantic Mapping
COP
Content ontology patterns (COPs)
Layer 1: Structured Heterogeneous Data
HL7 CDA
openEHR
ISO 13606
LOCAL SCHEMA

Figure 1. Semantic-driven Architecture

ing the same content: [*Disease*: *Inflammation*] and [*Body location*: *Myocardium*] vs. [*Disease*: *Myocarditis*] [3]. In the first example, the semantic relation between the disease and the body location is *not* provided by the clinical model. Medical data are provided at different level of detail, depending on the context requirements. E.g. smoking-related information recorded by a GP will be coarser than information recorded in a smoking cessation service. This implies co-existing clinical models, providing different information level detail. Contextual information is essential for an accurate processing (e.g. past or present history, etc.). The way information has been acquired (e.g. by machine, nurse, physician, patient) will determine its degree of trust. Each data element in a clinical model should provide this information in a standardized way. Some clinical data is coded using terminologies such as SNOMED CT, which are grounded on a model which explicitly formalizes the meaning of their terms. However, the selection of a term is usually driven by its textual description, which is often imprecise, and not by its formal definition. Additionally, some terms blend contextual with clinical information (boundary problem [4]).

Semantic mapping (Layer 2): Semantic content patterns are introduced to bridge between structured data and their semantic representation. They describe recurring information structures and provide a particular view on the underlying model of meaning, tailored to the needs of particular use cases [5], preventing users from a deep knowledge of the underlying formalisms. Patterns can be organized in hierarchies, in which top-level patterns can be specialized following a paradigm similar to object-oriented design, and in which their composition permits to cover larger modelling use cases. Patterns may be encoded as Subject-Predicate-Object (SPO) triples, to which data elements and value restrictions of clinical models are mapped. As a difference with the non-related data elements of clinical models, predicates constitute direct links between classes as dictated by the pattern. E.g. <*InformationItem* 'describes situation' *Inflammation*>, <*InformationItem* 'describes situation' *Disease*> and <*Disease* 'has location' *Myocardium*>.

Semantic mediator (Layer 3): The core is constituted by a set of ontologies which formalize clinical data meaning: an information entity ontology and a medical domain ontology, constrained by means of canonical dissections and relations provided by a top-level ontology. We consider SNOMED CT a good candidate as medical domain ontology, due to its high coverage and its grounding in formal-ontology principles [6]. A strict separation between *Information* and the entities represented by this information is fundamental. E.g. *Diagnosis* is a piece of information acquired by some clinician, at some place, at some point in time by following certain method. A *Diagnosis* represents a possibly clinical situation of the patient such as inflammatory disease. A cause of the *Inflammatory disorder* can be a *Viral infection*. *Severity* information diagnosed is, however inferred by the clinician and is a modifier of the information object. This separation is essential to use consistently information models and clinical terminologies.

Ontologies lend themselves to be expressed in logic, e.g. description logics (DL) [7], which enables inferencing, with two main goals: (1) detect semantic equivalences across different distributions of content between information models and terminologies; (2) advanced exploitation of clinical information by means of semantic queries. The first goal means that if the same data are represented by using a different combination of data elements and value restrictions, by inference it is possible to detect their equivalence (e.g. [*Disease*: *Inflammation*] and [*Body location*: *Myocardium*] **equivalent to** [*Disease*: *Myocarditis*]). The second goal refers to be able to query data at different level of detail (e.g. patients with inflammatory diseases, with heart disorders, etc.).

Virtual homogeneous data (Layer 4): It provides a homogeneous view on clinical data extracted from heterogeneous systems. Data are expressed at different detail levels, but they can be accessed homogeneously thanks to the underlying ontology-based annotations. As opposite to traditional databases where complete information is assumed, ontologies assume incomplete information, which affects how data can be queried. Query patterns based on the ontology content patterns used for the semantic mapping should guide queries.

Application (Layer 5): Here we find clinical systems and services with different information needs. A public health system might be interested in aggregating information results to extract population-level information while a primary care system might be interested in the results of a lab test. A decision support component might need to retrieve one very specific data value. Each use case requires querying data at different detail and precision. Besides, all the data retrieved might not deserve the same trust level and it must be explicitly shown (e.g. seniority of the author, supporting evidence, indication of whether a diagnosis is tentative or confirmed etc.).

2. Architectural Challenges

Structured heterogeneous data (Layer 1): Clinical data rendered according to clinical models or particular database schemas provide vague and mostly implicit contextual information. E.g. “93 kg” assigned to the data element “body weight” leaves open whether the weight is reported by the patient or has been measured in the clinic. Context is often expressed in natural language. In database schemas it is often not even documented. Therefore, the interoperability degree we can reach is directly dependent on the quality of the source *information* (i.e. data in context). The challenge is to optimize the amount of precisely expressed information quality for a reasonable effort.

The direct annotation of clinical data by ontology-based vocabularies like SNOMED CT may help clarify data meaning at a reasonable workload. However, the role of terminological resources in the EHR should be clear. Should they limit themselves to represent medical entities like “myocarditis”, or should they claim to provide codes even for complex statements such as “possible myocarditis” which combine a domain term with epistemic information about its specific use by the author of this statement. If the former case is preferred, then information models should represent the epistemic information. In case both variants are valid, interoperability requires additional effort (e.g. the code for “myocarditis” bound to an information model on diagnosis including diagnostic certainty attributes should be found to be equivalent to the code “possible myocarditis”). This binding has to be guided by the semantic categories of the underlying model of meaning and not primarily by its label, as this can be ambiguous. E.g. “myocarditis” could be interpreted as a disease or a morphology.

Semantic mapping (Layer 2): The main challenge in this layer is to use semantic content patterns to package information in order to provide a user friendly handling that facilitates unambiguous mapping of data elements and value sets. This requires that patterns provide enough expressivity to represent data and their rendering in a formal and interchangeable way, mainly hidden to the mapping builder, who should be supported by specific tools. Patterns encode repetitive modelling issues. They should be compositional, and they should be arranged in hierarchies, from top-level patterns (e.g. process plan) to specific ones (e.g. medication administration pattern). Their composition and specialization will be formalized by means of cardinality and value restrictions. As an example of composition, a comprehensive diagnosis pattern encodes a disease, together with its cause, location, etc. The latter two are provided by the situation (finding) pattern. Tools should also support the construction and management of patterns by a community of experts. Moreover, mechanisms to check the validity of clinical data with regards to patterns constraints should also be provided.

Semantic mediator (Layer 3): The main challenge is the model of meaning (ontology) itself. It must be expressive enough to address the complexity of representing clinical facts and doing it in a formal way, in order to be machine-processable. Besides, it must be independent of particular applications and provide relatively generic content that can be reused by heterogeneous systems. The use of a top-level ontology provides a set of canonical top-level classes and relationships that allow describing more complex content, using a set of generic and standardized language constructs, which facilitates the combination of different domain ontologies. However, although top-level ontologies can guide the construction of the model of meaning, distinguishing between what is information and what is not, or the so called boundary problem is not trivial.

The rendering of the model of meaning by using a logic-based language allows performing logical reasoning like inferring myocarditis from the combination of inflammation and myocardium, to cite a trivial example. However, reasoning on expressive ontologies is computationally expensive [8]. The size of data is also a challenge if semantic technologies are used [9]. Given the state of the art of semantic technologies, the biggest challenge consists in finding intermediate scalable solutions which leave the door open to include the upcoming progress in the field. The rendering of the model of meaning using a logic or non-logic based language, as well as the type of reasoning done, if any, will depend on the technological state of the art and the semantic interoperability requirements. Moreover, several renderings of a rich representation might be required, each attending a different purpose, e.g. data validation vs. data query.

Virtual homogeneous data (Layer 4): At this layer data can be accessed homogeneously and have – ideally – their complete meaning (ontological and epistemic) expressed in a formal language, from which logical entailments can be computed. If queries can be supported by a logical reasoner [10], answers might include derived facts. Scalability problems may have to be addressed by query rewriting, parallelization, optimization of reasoning algorithms, etc. For persisting semantically enriched data, native and non-native triple stores exist [11]. Native triple stores provide their own database implementation and non-native ones rely on particular relational databases. Native triple stores perform better since their performance has been optimized for that, however the support of reasoners is partly implemented and might be still expensive. Since ontologies assume incomplete information, if something has not explicitly negated it is not considered false as opposed to traditional databases. This has to be considered in the formulation of queries, which constitutes a barrier for most users. The use of query patterns based on principles similar to the content patterns described above could hide that complexity.

Application (Layer 5): Application systems at this level make queries based on the model of meaning, whose expressivity will determine that queries at different granularity level lead to answers. Since application systems might be placed within different medical contexts and be driven by different requirements and use cases, that data can be retrieved within their context is essential for their safe use. Close-to-user interfaces and the use of query templates facilitate query elaboration and result interpretation.

3. Conclusions

The described layered, semantic-driven architecture is centered around a semantic mediator, which is constituted by ontologies that formalize clinical data meaning. Their underlying logic allows for computing semantic equivalences across different distributions of content between information models and terminologies, and for advanced exploitation of clinical information by means of semantic query possibilities. Scalability problems are a known issues in logic-based models, therefore formalisms, not based on logic, can be considered depending on the particular purpose (e.g. data validation vs. data query). Semantic patterns offer a user-friendlier way to encode knowledge, which hides the complexity of the underlying model of meaning. They should be flexible and expressive enough to encode clinical models data but at the same time follow strict constraining principles. On top of that, their use by professionals requires of tools that support each of the tasks carried out at each layer of the architecture. E.g. Building and maintenance of patterns, mapping of data supported by patterns, query design, query interfaces, etc. The challenge now for the clinical / informatics communities is to grow libraries of such patterns, to help inform the design of future EHR repositories and message standards.

References

[1] V.M. Stroetmann et al. *Interoperability for better health and safer healthcare. Deployment and research roadmap for Europe*. ISBN-13 : 978-92-79-11139-6

[2] *SemanticHealthNet Network of Excellence*. http: //www.semantichealthnet.eu/

[3] K.E. Campbell, A.K. Das, M.A. Musen, A logical foundation for representation of clinical data, *JAMIA* **1** (3) (1994), 218-232.

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

[5] 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]

[6] S. Schulz, L. Jansen, Formal ontologies in biomedical knowledge representation. *Yearbook of Medical Informatics* **8** (2013), 132-146.

[7] F. Baader et al. (eds.) *The Description Logic Handbook. Theory, Implementation, and Applications* (2nd Edition). Cambridge University Press, 2007.

[8] I. Horrocks, Reasoning with Expressive Description Logics: Theory and Practice, Lecture Notes in Computer Science **2392** (2002), 1-15.

[9] R. Möller, C. Neuenstadt, Ö. L. Özçep, S. Wandelt, Advances in Accessing Big Data with Expressive Ontologies, *Lecture Notes in Computer Science* **8077** (2013), 118-129.

[10] I. Kollia, B. Glimm, I. Horrocks. SPARQL Query Answering over OWL Ontologies. *Proceedings of the 8th Extended Semantic Web Conference* 2011, 382-396.

[11] Comparison of Triple stores: http://www.bioontology.org/wiki/images/6/6a/Triple_Stores.pdf