Health Informatics Meets eHealth
G. Schreier et al. (Eds.)

doi:10.3233/978-1-61499-645-3-63

Automated Transformation of openEHR Data Instances to OWL

Birger HAARBRANDT[a,1], Thomas JACK[b], Michael MARSCHOLLEK[a]
[a]*Peter L. Reichertz Institute for Medical Informatics, University of Braunschweig and Hannover Medical School, Germany*
[b]*Department of Pediatric Cardiology and Intensive Care Medicine, Hannover Medical School*

Abstract. Standard-based integration and semantic enrichment of clinical data originating from electronic medical records has shown to be critical to enable secondary use. To facilitate the utilization of semantic technologies on clinical data, we introduce a methodology to enable automated transformation of openEHR-based data to Web Ontology Language (OWL) individuals. To test the correctness of the implementation, de-identified data of 229 patients of the pediatric intensive care unit of Hannover Medical School has been transformed into 2.983.436 individuals. Querying of the resulting ontology for symptoms of the systemic inflammatory response syndrome (SIRS) yielded the same result set as a SQL query on an openEHR-based clinical data repository.

Keywords. Electronic Health Records*/standards, openEHR, Biomedical Ontology

1. Introduction

Standard-based data integration is one of the premises for leveraging clinical data for secondary use. This requirement has driven the development of a set of methods that are often referred to as detailed clinical models (DCM) [1]. DCM aim for unambiguous, formalized, computable and manageable representation of domain content models. The evolving two-level modelling approach utilized by openEHR is layered as Archetype Model (AM) and Reference Model (RM) [2]. DCM can be complemented by the use of standardized medical terminologies. Comprehensive nomenclatures like SNOMED CT[2] address the need for standardized and unambiguous identifiers of clinical concepts, but also encompass ontological characteristics [3]. These characteristics allow utilization of semantic technologies to infer knowledge from medical data, merge data from diverse repositories, to calculate semantic similarities between patients and to identify cohorts for clinical research. Additionally, first studies have demonstrated the combined use of semantic web representations of archetypes and the Semantic Web Rule Language[3] (SWRL) to support clinical decision support and the calculation of quality indicators [4][5]. However, mapping of clinical data from source systems had to been done manually in these studies due to proprietary and non-standardized data structures and information models. With this work, we wanted to investigate if clinical data that is based

[1] Corresponding Author: Birger Haarbrandt, Peter L. Reichertz Institute for Medical Informatics, University of Braunschweig and Hannover Medical School, Carl-Neuberg-Strasse 1, 30625 Hannover, Germany, E-Mail: birger.haarbrandt@plri.de

[2] http://www.ihtsdo.org/snomed-ct

[3] https://www.w3.org/Submission/SWRL/

on the openEHR Reference Model, can be automatically transformed into instances of corresponding Web Ontology Language [4] (OWL) representations of archetypes to facilitate the secondary use of this data.

2. Methods

As a proof-of-concept study, we developed a software program to automatically transform real-world clinical data from an openEHR-based clinical data repository (CDR) into OWL individuals. The openEHR specification describes an open, interoperable electronic health record, built upon the two-level modelling approach [2]. By offering flexible standard-based data integration, the use of openEHR allows the development of systems that are to a certain degree independent from the clinical data models which are expressed as archetypes using the Archetype Definition Language (ADL). Each archetype represents a comprehensive collection of attributes, forming a maximum data set, to be able to address the requirements for medical documentation of any thinkable use-case [2].

The CDR is part of the Hannover Medical School Translational Research Framework (HaMSTR), a research data warehouse to investigate the use of DCM for secondary use. HaMSTR follows an Inmon architecture (i.e. data is normalized in an enterprise data warehouse before being delivered to customers through data marts). It contains data from two patient data management systems and the electronic medical record. All clinical data stored in the database is represented by serialized objects of the openEHR reference model. To store openEHR-based data, it uses a hybrid approach, utilizing both relations and *Extensible Markup Language* (XML) documents for persistence. HaMSTR uses openEHR compositions as data input. Compositions can be thought of as clinical documents composed of a set of archetypes to cover specific clinical use cases. They are represented in openEHR XML, which is a commonly used representation of serialized instances of the openEHR Reference Model. These XML documents are analysed and sections of particular archetypes are extracted and stored in corresponding tables having columns with xml data fields and an ID. This approach preserves the hierarchical structures of the templates while allowing a certain degree of optimization on the column level.

To build an openEHR reference model to OWL converter (RM2OWL), we utilized and enhanced the work previously done by Lezcano et al. [4], who developed the ADL2OWL translator and introduced the ArchOnt framework. The latter is the conception of a toolchain to allow the semantic enrichment of clinical data. While the ADL2OWL translator permits the automated translation of openEHR archetypes into OWL, it does not address the need for *automated* transformation of real-world clinical data into adequate OWL individuals.

The RM2OWL converter and involved parts of HaMSTR have been implemented using Java 8, SQLite and a Microsoft SQL Server 2012 database. To build the software, we utilized rapid prototyping development methodology. The libraries that were mainly used are the OWLAPI[5] and the Java Reference Implementation of openEHR[6].

[4] https://www.w3.org/TR/owl2-overview/

[5] http://owlapi.sourceforge.net/

[6] https://github.com/openEHR/java-libs

To test the correctness of the transformation, we transformed and loaded de-identified data from the CDR of HaMSTR into a triple store (*Stardog 4.0.3*[7]) and used SPARQL to query for paediatric patients matching symptoms of systemic inflammatory response syndrome (SIRS).

3. Results

Figure 1 shows the basic components of the RM2OWL converter and their role in the transformation and semantic enrichment process. Basically, the architecture is a implementation of the proposed workflow of the ArchOnt framework [4]. We slightly modified the ADL2OWL translator to introduce consistent naming convention between archetypes and their OWL representations, to check created ontologies for consistency and satisfiability (using the pellet reasoner[8]) and to make it part of a newly developed *Knowledge Store*. The *Knowledge Store* (based on a SQLite database) stores all archetypes used in the clinical data repository in both, ADL and OWL. Whenever an archetype is added to the store, a corresponding OWL representation gets automatically created. Our RM2OWL converter introduces functions to (1) query the CDR for archetype qualifiers, (2) obtain and merge OWL representations of archetypes from the *Knowledge Store*, (3) resolve terminology bindings to establish relations to terminologies/ontologies for semantic enrichment, (4) automatically transform openEHR data to OWL individuals, and (5) export the result to a triple store.

The first step of each transformation is the provision of a set of encounter numbers as input parameters. Optionally, a list of archetype qualifiers can be stated to limit the transformation to a defined subset of clinical concepts. Following, the converter generates a list of all archetypes that have been incorporated in matching templates. The list is used to retrieve respective ADL und OWL representations of archetypes from the *Knowledge Store*. Then, the OWL representations are merged into a single ontology and automatically checked for consistency (again using pellet).

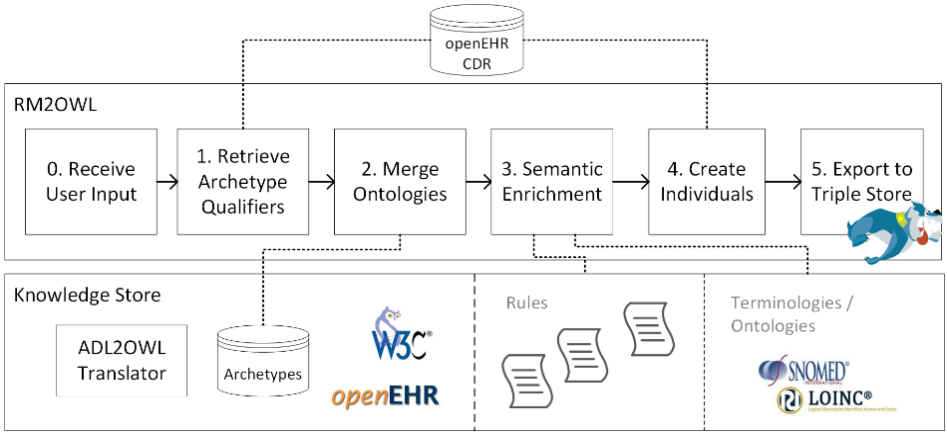

Figure 1: Overview of the transformation steps of the RM2OWL converter and the involved components.

[7] http://stardog.com/

[8] https://github.com/clarkparsia/pellet

Additionally, for *semantic enrichment* of the data, the ontology section of each archetype is inspected for available terminology bindings. If a binding exists and the terminology is available in the *Knowledge Store,* the system adds an OWL *equivalent class* axiom to the ontology and subsequently merges the archetype ontology with the target terminology.

After building the ontology, the system is set to *create individuals* from openEHR instance data. The RM2OWL converter retrieves all matching XML documents that are associated with the given encounter numbers. The transformation is done template-wise to be able to preserve hierarchical relationships that have been defined in a template. As the transformation of templates is not yet support by the ADL2OWL translator, we circumvent this shortcoming by only creating individuals of the generic *Composition* concept with a name property holding the template's name. Additionally, we added concepts to the ontology to conveniently represent subject IDs and encounters.

The transformation of each XML document is done by depth-first search using the Document Object Model representation. Generally, the transformation is rule based, i.e. each structure within the XML document triggers a function to create individuals of corresponding concepts. All information required to create an OWL individual is retrieved from the corresponding AM by using the *archetype_node_id*. Through applying the same naming convention as the ADL2OWL Translator, including a newly

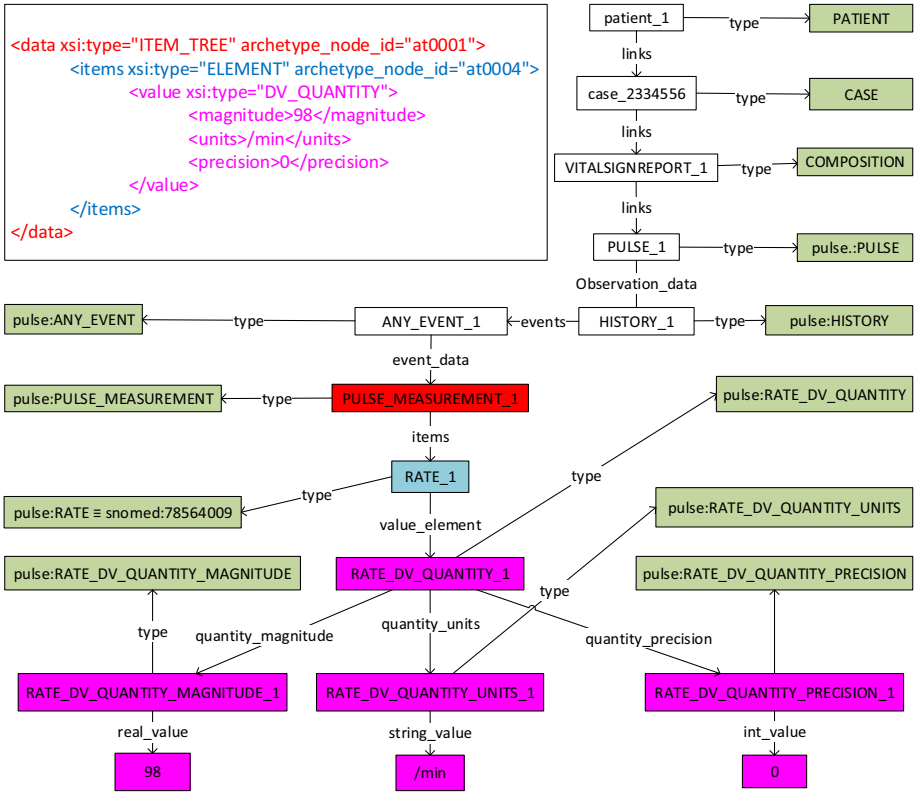

Figure 2: Example of an openEHR instance data transformation. Green indicates concepts originating from the OWL representations of the pulse archetype and concepts of the ontology. The colors used in the code listing (top left) and the elements indicate which individuals result from which element in the openEHR XML file.

added namespace based on the qualifier of the respective archetype, an individual can unambiguously be created and assigned to a concept. In order to establish object properties between individuals, a reference of each parent individual is passed to the function that is responsible to create a child. As every data point in openEHR XML is encapsulated by a defined data type of the reference model, it is possible to automatically generate literals for each value at the document's leaf nodes.

Figure 2 illustrates the interrelationship between instances of clinical data represented in openEHR XML, the OWL representation of archetypes and created individuals on the example of a pulse archetype. To simplify and make the figure more understandable to readers, some elements have been omitted. The colours indicate how a fragment looks before and after the transformation.

The example starts after the creation of individuals representing the patient, an encounter number and a stub for the template has been completed. To represent relationships between archetypes and templates, we use the *links* object property as proposed by Dentler et al. [5]. At first, the *data* node within the openEHR XML document is reached. It is checked for its type attribute, which in this case is an *ITEM_TREE*. With the help of the *archetype_node_id* ("at0001") the converter is able to perform a lookup of the element's name defined in the ontology section of the archetype. In the case of the pulse archetype, the ITEM_TREE has been constrained to represent a *PULSE_MEASUREMENT* (i.e. a tree structure holding data and state information of a single measurement of a pulse).

In the next step, an items element of type ELEMENT is found as a child of the data node. Through the *archetype_node_id* ("at0004") it can be identified as the RATE element of the archetype which contains the rate of a pulse, measured in beats per minute. As the measurement of a pulse rate results in a numeric value, it is represented by the DV_QUANTITY data type, enclosing data fields for magnitude, precision and units. The corresponding values get extracted using XPath expressions and are used to create OWL literals.

The binding between concepts in the archetype ontology and a terminology (e.g. SNOMED CT) is shown on the example of the Rate. An equivalent class axiom binding the concept of the OWL archetype and the concept having the SNOMED CT ID 78564009 has been introduced during the *semantic enrichment*. This binding allows queries and rules to be applied across merged ontologies.

Finally, the created ontology including archetypes, individuals and terminologies is loaded into a triple store (e.g. Stardog 4.0.3). By applying this approach, we have been able to successfully create OWL representations and individuals for all observational archetypes from the Knowledge Store. Table 1 provides a list of all incorporated archetypes.

Table 1: overview of archetypes used in HaMSTR and tested for transformation

Archetype	
OBSERVATION.blood_pressure.v1	OBSERVATION.respiration.v1
OBSERVATION.body_temperature.v1	OBSERVATION.body_weight.v1
OBSERVATION.pulse.v1	OBSERVATION.indirect_oximetry.v1
OBSERVATION.glasgow_coma_scale.v1	OBSERVATION.height.v1
OBSERVATION.braden_q_scale.v1	EVALUATION.problem_diagnosis.v1
OBSERVATION.lab_test.v1	EVALUATION.health_risk.v1

4. Evaluation

To test the correctness of the RM2OWL converter, we transformed and queried data of a subset of patients from the pediatric intensive care unit of Hanover Medical School. As an early step to develop a decision support system that alerts physicians about suspected presence of the systemic inflammatory response syndrome (SIRS), we conducted a retrospective cohort identification of patients matching the defined criteria for SIRS. Early detection of SIRS is of high clinical relevance, as it is closely related to sepsis (defined as SIRS with proven infection), organ dysfunction and organ failure which are among the most common reasons for morbidity and mortality in pediatric intensive care medicine [6].

In pediatric patients, age-specific vital signs and laboratory values are used for clinical diagnosis of SIRS. As SIRS criteria were not derived by population studies but defined by experts at the International Consensus Conference on Pediatric Sepsis (IPSCC), there is a lack of evidence for the exact ranges of vital signs for the different ages [6]. Two or more of the following values need to be present (of which one must be an abnormal temperature or leukocyte count): (1) fever (>38.5°C) or hypothermia (<36°C), (2) tachycardia, (3) tachypnea or (4) an abnormal blood leucocyte count. Except for the body temperature, all ranges of the variables depend on the patient's age. The definition of the IPSCC divides pediatric patients into six different age groups (<1 week, 1 week to 1 month, 1 month to 1 year, 2-5 years, 6-12 years, 13-<18 years).

We took a sample of 229 patients, which is all patients that had an encounter at the pediatric ICU in the first quarter of 2015 (01.01.2015 – 31.03.2015). All data has been de-identified; therefore, no demographic information has been available for querying. Hence, we applied the least rigid criteria for adolescent patients (13-<18 years) in order to retrieve a subset with a high recall and low precision for further processing.

```
PREFIX bt:<http://mh-hannover.de/openEHR-EHR-OBSERVATION.body_temperature.v1#>
PREFIX pulse: <http://mh-hannover.de/openEHR-EHR-OBSERVATION.pulse.v1#>
PREFIX resp: <http://mh-hannover.de /openEHR-EHR-OBSERVATION.respiration.v1#>
PREFIX oer: <http://klt.inf.um.es/~cati/ontologies/OpenEHR-SP-v2.0.owl#>
PREFIX lab: <http://mh-hannover.de /openEHR-EHR-OBSERVATION.lab_test.v1#>
PREFIX patient: <http://klt.inf.um.es/~cati/ontologies>

SELECT DISTINCT ?indiv  WHERE
  {
 …
 OPTIONAL{
 ?indiv oer:has/cati:observation_data/oer:events ?event .
 ?event a lab:ANY_EVENT .
 ?event oer:event_data/oeri:items/oer:value_element/oer:quantity_magnitude/oer:real_value ?lab .
 ?event oer:event_data/cati:items/oer:value_element/oer:value_text/oer:string_value ?testName .
 FILTER ((?testName = 'Leukozyten'^^xsd:string)
 } BIND(MIN(?lab) AS ?minLab)

FILTER ((?maxLab >= 11.0 || ?minLab <= 4.5) || (?maxBT >= 38.5 || ?minBT <= 36.0)) .
  FILTER ( ((?maxLab >= 11.0 || ?minLab <= 4.5) && (?maxBT >= 38.5 || ?minBT<= 36.0)) ||
((?maxLab >= 11.0 || ?minLab <= 4.5) && (?maxPuls >= 110)) || ((?maxLab >= 11.0 || ?minLab <=
4.5) && (?maxResp >= 20)) || ((?maxBT>= 38.5 || ?minBT <= 36.0) && (?maxPuls >= 110)) ||
((?maxBT >= 38.0 || ?minBT<= 36.0) && (?maxResp >= 20)) || ((?maxPuls >= 110) && (?maxResp
>= 20)) )
   }
 }
```

Figure 3: Excerpt of a SPARQL query to identify patients suspected of SIRS.

In sum, the converter created 2.983.436 individuals. To test the accurate transformation of the data, we performed semantically identical queries using both, SQL and SPARQL. The SQL query was conducted on the relational tables of respective data in HaMSTR. Figure 3 presents an excerpt of the SPARQL query, showing the pattern to retrieve the leucocyte count from the OWL representation of the *lab_value* archetype and the filter statement to retrieve only patients for which at least two criteria are matching. In total, the query returned the surrogate IDs of 153 patients that have met the stated inclusion criteria. We were able to retrieve the exact same result set as by using an equivalent SQL statement on the relational database of HaMSTR.

5. Discussion

Our work contributes a novel method to enable automated transformation of clinical data into OWL individuals. The converter uses a generic and fully automated transformation of openEHR instance data which exploits the strength of two-level modelling: changing requirements of an application system can be addressed by the definition or the versioning of archetypes without making alterations to the underlying reference model. Therefore, we are optimistic that the source code of the converter will not need to be modified when new clinical concepts are integrated in the CDR (or any other openEHR-based application system). By enabling such a generic transformation, the use of openEHR to represent clinical data can help to save resources for maintenance and to allow fast and correct provision of OWL representations of patient data to researches. Thereby, the RM2OWL converter might help to promote research on semantic technologies in real-world clinical environments. For example, investigation on automated semantic consistency checking and case-based semantic similarity measurements can be conducted more efficiently when customized transformation scripts from source data to OWL instances are not needed anymore.

Moreover, even in cases where these advanced research questions are not targeted, the representation of openEHR instance data in OWL can complement data retrieval based on SQL or the Archetype Query Language (AQL) by allowing semantic queries using SPARQL. For example, while AQL currently lacks operators to join data from diverse templates, this is possible with SPARQL [9].

Related work has been conducted in the SHARPn project [7] (using the Clinical Element Models (CEMs) approach instead of openEHR) and by Fernández-Breis et al. [8]. The former describes the transformation of the CEMs (which can be compared to archetypes) to OWL representations. However, there is a lack of explanation if and how instance data is actually transformed to individuals. As openEHR and CEMs developed independently but are quite similar instances for two-level modelling approaches, our methodology might in principle be applicable for CEM instances as well.

Fernández-Breis et al. have demonstrated a (semi-)automatic transformation within a rich framework for utilization of clinical data for secondary use employing a hybrid approach that clearly separates data level and clinical knowledge by linking XML and OWL representations. In contrast, our transformation of data instances, based on the archetype transformation of Lezcano et al. [4], expresses all types of clinical data as individuals of OWL representations of archetypes. While this might limit the use of certain aggregation functions (though, SPARQL 1.1 shows improvement regarding this

matter), it allows exploiting the ontological characteristics of the openEHR reference model within semantic queries.

Limitations of this work exist regarding the transformation of *action* and *instruction* archetypes. At the time of this study, the CDR of HaMSTR only contained data according to *evaluation* and *observation* archetypes. Hence, we decided to develop the first version without the above mentioned entry level classes. Moreover, the representation of templates is not yet supported. Therefore, some metadata elements as the healthcare facility or participations of clinical personnel were not available for querying.

Further development of the converter will concentrate on three aspects: First, the transformation of templates and further entry classes needs to be implemented in the ADL2OWL translator to represent the full set of metadata represented in document headers. Second, the introduction of a more granular way of pre-selecting particular data elements from templates is needed to avoid the transformation of unwanted data. For example, if only the result of a laboratory measurement (represented by a lab value archetype) is needed for research, only this element should be transformed. Third, the creation of a graphical user interface to select the above mentioned elements and to help enter the parameters would make the software more easy to use. Simultaneously, there will be investigations addressing the need to learn about the practical implications of this approach for secondary use.

References

[1] Goossen W, Goossen-Baremans A, van der Zel M. Detailed clinical models: a review. Healthc Inform Res. 2010 Dec;16(4):201-14. doi: 10.4258/hir.2010.16.4.201. Epub 2010 Dec 31.

[2] Beale T. Archetypes, constraint-based domain models for future-proof information systems. In: Eleventh OOPSLA Workshop on Behavioral Semantics: Serving the Customer (Seattle, Washington, USA, Nov 4, 2002). 2002, pp. 16-32.

[3] Rector AL, Brand S. Why do it the hard way? The case for an expressive description logic for SNOMED. J am Med Inform Assoc. 2008 Nov-Dec15(6):744-51.

[4] Lezcano, L., Sicilia, M. A., & Rodríguez-Solano, C. (2011). Integrating reasoning and clinical archetypes using OWL ontologies and SWRL rules. Journal of biomedical informatics, 44(2), 343-353.

[5] Dentler K, ten Teije A, Cornet R, de Keizer N. Semantic Integration of Patient Data and Quality Indicators based on openEHR Archetypes. ProHealth 2012/KR4HC 2012, LNAI 7738:85–97, 2013.

[6] Goldstein B1, Giroir B, Randolph A. International pediatric sepsis consensus conference: definitions for sepsis and organ dysfunction in pediatrics. Pediatr Crit Care Med. 2005 Jan;6(1):2-8.

[7] Fernández-Breis JT1, Maldonado JA, Marcos M, Legaz-García Mdel C, Moner D, Torres-Sospedra J, Esteban-Gil A, Martínez-Salvador B, Robles M. Leveraging electronic healthcare record standards and semantic web technologies for the identification of patient cohorts. J Am Med Inform Assoc. 2013 Dec;20(e2):e288-96.

[8] Tao C1, Jiang G, Oniki TA, Freimuth RR, Zhu Q, Sharma D, Pathak J, Huff SM, Chute CG. A semantic-web oriented representation of the clinical element model for secondary use of electronic health records data. J Am Med Inform Assoc. 2013 May 1;20(3):554-62.

[9] openEHR Foundation. Archetype Definition Language ADL2. Available from: http://www.openehr.org/releases/1.0.2/architecture/am/adl2.pdf [last accessed January 26, 2016].