Journal of Biomedical Informatics xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Semantic processing of EHR Data for clinical research

Hong Sun *, Kristof Depraetere, Jos De Roo, Giovanni Mels, Boris De Vloed, Marc Twagirumukiza, Dirk Colaert

Advanced Clinical Applications Research Group, Agfa HealthCare, 100 Moutstraat, Gent 9000, Belgium

ARTICLE INFO

Article history:
Received 9 May 2015
Revised 10 September 2015
Accepted 17 October 2015
Available online xxxx

Keywords:
Semantic interoperability
N3 rules
EHR
RESTful
Clinical research
Semantic web stack

ABSTRACT

There is a growing need to semantically process and integrate clinical data from different sources for clinical research. This paper presents an approach to integrate EHRs from heterogeneous resources and generate integrated data in different data formats or semantics to support various clinical research applications. The proposed approach builds semantic data virtualization layers on top of data sources, which generate data in the requested semantics or formats on demand. This approach avoids upfront dumping to and synchronizing of the data with various representations. Data from different EHR systems are first mapped to RDF data with source semantics, and then converted to representations with harmonized domain semantics where domain ontologies and terminologies are used to improve reusability. It is also possible to further convert data to application semantics and store the converted results in clinical research databases, e.g. i2b2, OMOP, to support different clinical research settings. Semantic conversions between different representations are explicitly expressed using N3 rules and executed by an N3 Reasoner (EYE), which can also generate proofs of the conversion processes. The solution presented in this paper has been applied to real-world applications that process large scale EHR data.

1. Introduction

After decades of development of electronic health records the integration of health records from different EHR systems has become a rising demand. The development of standard clinical information models is an attempt to tackle the storage and exchange of clinical data. Standards like HL7 [1], openEHR [2], ISO 13606 [3], etc. are developed to store or exchange patient records with structured formats. However, the semantic interoperability between different standards remains a challenge.

In order to cope with semantic differences between EHRs we mapped data from different systems to semantic representations expressed with a core ontology [4] in the DebugIT project [5]. Clinical data stored in different EHR systems are first mapped to semantic representations with local semantics of their respective EHR system and then mapped to expressions using the core ontology [6].

To further improve reusability, in the later SALUS project [7], we built semantic patterns for relevant clinical domains [8,9] by reusing existing public ontologies and standard terminologies. We named such patterns Clinical Research Entity Advanced Model (CREAM) and aimed to achieve semantic interoperability between EHR systems and clinical applications by mapping their data to semantic expressions following CREAM.

The semantic interoperability achieved through a core ontology, or a harmonized domain information model such as CREAM, is nevertheless still fragile: interoperability is only achieved when the involved parties make their commitment to the common information model. Stakeholders of this common information model are not limited to the data providers (i.e. EHR systems) but also include data consumers (e.g. many clinical research applications). It is difficult to adapt research applications, which are built on top of a dedicated clinical data model, e.g. i2b2 [10], OMOP [11], to directly consume data expressed with domain semantics such as CREAM.

It is therefore important that data from a clinical data source can be mapped to a set of representations so as to achieve interoperability between the data source and multiple clinical research applications. This paper introduces a semantic data virtualization (SDV) scheme which is able to build multiple semantic data virtualization layers on top of a data source. Thus multiple clinical research applications can be supported. The SDV generates data in requested formats or semantics on demand. There is no need to dump the data to various representations upfront, thus avoiding the burden of synchronization with the source.

The SDV is constructed using RESTful services, which enables data transformation in a fully automated way. We use N3 rules to create the mappings between graph patterns expressed with

* Corresponding author. Tel.: +32 3444 8108.
E-mail address: hong.sun@agfa.com (H. Sun).

http://dx.doi.org/10.1016/j.jbi.2015.10.009
1532-0464/

different ontologies in different domains. As most of the existing N3 reasoners, e.g. EYE [12], CWM [13], etc. can generate a proof of a conversion process, we not only express the semantic conversions in an explicit way, but also the proof explains the executed conversion process. The SDV presented in this paper has been applied in the SALUS project to build the semantic interoperability layer. This paper generalizes the software components contained in the SDV from their specific implementation of the SALUS project, so that the SDV can be used in other projects.

The rest of this paper starts with discussions of related work. A summary of the different data layers in semantic processing of EHR for clinical research is presented, followed by an introduction of the architecture of the SDV and examples of mapping data between different layers by the SDV. Applications of the SDV in the SALUS and the AP-HP project are also given, together with a brief discussion regarding its performance.

2. Related work

To improve the interoperability of EHRs represented with different standards, mappings between different standards are developed and domain ontologies are created. Costa et al. [14] developed source ontologies for openEHR and ISO 13606, as well as a domain ontology which bridges the two standards. Data transformation is carried out through syntactic mapping between the archetypes of openEHR, ISO 13606 and the archetype model of the domain ontology. Gonçalves et al. [15] build up a domain ontology of ECG and map schemas of three ECG standards directly to the domain ontology. They find it difficult to develop a domain ontology to which different standards can directly be mapped.

Besides the report of directly mapping source data to a semantic representation with domain ontologies there are also proposals to achieve the mapping through multiple steps, which first map the EHR data with source semantics and later convert to representations with domain ontologies. Martínez-Costa et al. [16] describe the data layers between data sources and end applications and state the necessary steps towards EHR semantic interoperability, as well as the challenges in implementing the steps. Berges et al. [17] first obtain the ontological representations of relational databases and later map the database ontologies to their canonical ontology. The canonical ontology presented in [17] reuses existing medical terminologies such as LOINC and SNOMED.

Different methods are proposed to represent the target clinical model. Early research relies on using a core ontology to represent the target clinical model [5,14,15]. However, the ontology itself does not provide guidelines on how it can be used to represent target clinical models. Ontology content patterns, which guide and standardize the meaning of the content of clinical models, are proposed as a close-to-user representation to improve reusability [8,18,19]. The Clinical Information Modeling Initiative (CIMI) [20] proposes a set of modeling patterns, defined as clinical models, that can act as guidelines for the creation of ontology content patterns. The National Patient-Centered Clinical Research Network (PCORnet) also developed their Common Data Model (CDM) [21] to map data from PCORnet partners to a common model.

Many of the above mentioned approaches have been applied in projects that target semantically processing of EHR data for clinical research. The DebugIT project [5] maps EHR data from eight hospitals across Europe to representations with a core ontology for epidemiological research. Clinical questions are expressed with the DebugIT core ontology. A query generation service translates the clinical questions to corresponding SPARQL queries expressed with the source ontology, which are then executed on SPARQL endpoints at each site. A conversion service using an EYE reasoner maps the source data to expressions with the core ontology. The EHR data remain at the local hospitals and the outcomes of clinical questions are aggregated and displayed in the central dashboard.

The Strategic Health IT Advanced Research Projects (SHARPn) [22] develops their Clinical Entity Models (CEMs) as target model for EHR processing. Natural language processing (NLP) is used to process unstructured data. EHRs from two data sources, after NLP and normalization processing, are mapped to XML instances that conform to the CEM XSDs and stored in a central repository after anonymization. In their use case EHR data from 10,000 patients are used for diabetes mellitus research [23].

The EHR4CR project [24] aims at reusing EHR data for clinical research purposes. Two clinical data warehouses (CDW) are used as target models: an i2b2 data warehouse and a data warehouse with an EHR4CR specific schema. In their pilot application [25] each site established a CDW locally and used an ETL process to load the CDW with data from their respective EHR system. Seven sites use the EHR4CR database schema for their CDW, while the remaining four sites use i2b2. They are able to send one centrally created feasibility query and execute it at eleven sites to receive aggregated feasibility numbers with two different types of database schemas. Since both the SHARPn and the EHR4CR project store converted data in separate data warehouses, they have the burden to keep the data synchronized between the data source and their target data warehouse.

The SALUS project [7] aims to create the necessary semantic interoperability infrastructure to enable secondary use of electronic health records by various clinical tools for proactive post market safety studies. The semantic interoperability layer of the SALUS project is constructed following the semantic processing scheme presented in this paper. It uses a set of semantic patterns, namely CREAM, to represent the target clinical model. The source data to CREAM conversion is carried out on demand at run-time, which avoids maintaining extra data stores for converted data. The pilot application is carried out on two EHR systems which contain 1 million and 10 million patients respectively. Six clinical applications developed by different partners are successfully executed on the converted data at both sites.

3. Semantic data virtualization

The research presented in this paper aims to achieve semantic interoperability between data sources and clinical research applications through a data virtualization mechanism. This section first shows the data layers as well as the data flow in the proposed semantic framework. Then the architecture of the SDV is introduced. Examples of using the RESTful services of the SDV to implement the data flow are also demonstrated.

3.1. Data layers

> "*The purpose of abstraction is not to be vague, but to create a new semantic level in which one can be absolutely precise.*"
>
> [Edsger Dijkstra]

In our previous work we used a two-step formalization approach, which first formalized operational data with its local semantics, and later converted the data with local semantics to data with domain semantics [6]. Both local and domain semantics are precisely defined with ontologies. In this paper we extend this approach to further semantically process the data with domain semantics to application semantics, so as to achieve interoperability towards different clinical research applications. Fig. 1 shows the data layers in semantic data processing for clinical research as well as the needed actions to transfer data between the layers.

Layer 1: Heterogeneous source non-RDF data. On data layer 1 EHRs represented in different non-RDF formats are kept in their

Please cite this article in press as: H. Sun et al., Semantic processing of EHR Data for clinical research, J Biomed Inform (2015), http://dx.doi.org/10.1016/j.jbi.2015.10.009