EHR4CR: A semantic web based interoperability approach for reusing electronic healthcare records in protocol feasibility studies

Sajjad Hussain[1], David Ouagne[1], Eric Sadou[1], Thierry Dart[1], Marie-Christine Jaulent[1], Boris De Vloed[2], Dirk Colaert[2], Christel Daniel[1]

[1]INSERM, UMR_S 872, Eq. 20, Centre de Recherche des Cordeliers, Paris, France
[2]Advanced Clinical Application Group AGFA Healthcare, Belgium
{sajjad.hussain, christel.daniel, david.ouagne, eric.sadou, thierry.dart, marie-christine.jaulent}@crc.jussieu.fr; {boris.devloed, dirk.colaert}@agfa.com

Abstract. A major barrier to repurposing routinely collected data for clinical research is the heterogeneity of healthcare information systems. Electronic Health Record for Clinical Research (EHR4CR) is a European project designed to improve the efficiency of conducting clinical trials. We propose an initial architecture of the EHR4CR Semantic Interoperability Framework using Semantic Web technologies. We used a model-driven engineering approach to build the semantic resources and derived an HL7-based EHR4CR information model. We plan to apply existing ontology modularization techniques for obtaining relevant set of common data elements and clinical terminologies bound to our ERH4CR information model. EHR4CR platform provides semantic interoperability services that facilitate the process of defining eligibility criteria based on standard terminologies and then transforming and executing the user-defined SPARQL queries on RDF triplestore representing local EHRs. We plan to evaluate our semantic interoperability framework for patient eligibility determination in the context of 10 clinical trials running in 11 health institutions.

Keywords: Clinical research, semantic interoperability, electronic health record, clinical data warehouses, eligibility criteria, controlled vocabulary

1 Introduction

A major barrier to repurposing clinical data directly from Electronic Health Records (EHRs) or from Clinical Data Warehouses (CDWs) during clinical trial design and execution is that information systems in both domains – patient care and clinical research – use different information representations and terminology systems [1]. The collective efforts of multiple organizations (such as ISO, HL7, CDISC, etc) are currently focused on defining various standards required to achieve semantic interoperability and bridge the gap between clinical research and patient care by defining mappings between their data schemes [2]. However, the current scope of published mappings is limited. In the context of patient eligibility, due to the emerging requirements for representing new eligibility criteria from different domain areas, new mappings are needed [3]. We argue that integrating patient care and clinical research domains requires a standard-based and scalable semantic interoperability framework, allowing dynamic mappings between data structures and semantics of varying data sources.

The EHR4CR (Electronic Health Records for Clinical Research) project aims to improve the efficiency and reduce the cost of conducting clinical trials, through better leveraging of routinely collected clinical data in the trial design and execution life-cycle [4]. In this paper, we propose an initial architecture of the EHR4CR Semantic Interoperability Framework using Semantic Web technologies. EHR4CR platform provides semantic interoperability services that facilitate the process of defining eligibility criteria based on standard terminologies and then transforming and executing the user-defined SPARQL queries on RDF triplestore representing local EHR/CDWs.

2 Related Work: Bridging standards and semantic interoperability approaches

There have been various attempts in establishing semantic interoperability, developing different approaches for bridging clinical care and clinical research [5-8]. In this regard, one of the prominent attempts is the Shared Health Research Information Network (SHRINE), offering the SHRINE Ontology Mapping Tool (SHRIMP) for creating association mappings between local and standard vocabularies, and allowing researchers to define distributed queries over a federated network of i2b2 systems across multiple institutions [5]. The PONTE project focuses on developing advanced tools for semantic interoperability between clinical research and EHR data, which takes into consideration all widely used standards, such as coding systems, terminologies, vocabularies as well as health messaging standards such as HL7 [6]. The Linked2Safety project aims, similar objectives being targeted in EHR4CR, to advance clinical practice and accelerate medical research, by providing pharmaceutical companies, healthcare professionals and patients an innovative semantic interoperability framework facilitating the efficient and homogenized access to distributed EHRs [7]. A recently launched EURECA project aims to enable seamless, secure, scalable and consistent linkage of healthcare information residing in EHR systems with information in clinical research information systems [8].

Current interoperability efforts in both patient care and clinical research include defining metadata and vocabulary standards and using these to define Common Data Elements (CDEs) [9]. The CDISC SHARE project aims at building a metadata repository of CDEs based on the BRIDG model and its underlying mapping to the HL7 RIM. Another initiative is the Ontology of Clinical Research (OCRe), which provides methods for binding external information standards (e.g. BRIDG) and clinical terminologies (e.g. SNOMED CT) [9]. A recently launched initiative, Fast Health Interoperability Resources (FHIR) [10], offers the use of RDF-modeled core FHIR resources, and provides a semantic harmonization framework for establishing semantic interoperability between clinical research and patient care domains. To demonstrate the value of Semantic Web (SW) specifications in bridging patient care and clinical research, a

W3C task on Clinical Observations Interoperability was established [11]. As a use case for secondary use of EHRs for clinical research, checking clinical trial eligibility for patient recruitment was chosen and a prototype implementation was achieved. In line with this approach, we propose EHR4CR semantic interoperability framework based on SW technologies, where in our case, we define the mappings between standard clinical terminologies and local EHR terminologies, and use an HL7 based information model as a standard interface for querying heterogeneous CDWs.

3 EHR4CR Semantic Interoperability Framework

We propose the EHR4CR Semantic Interoperability Framework for consistent interpretation of clinical data accessed from varying sources (see Figure 1). A template-based query interface at the *User Workbench* allows clinical researchers to define eligibility criteria based on standard terminologies, data elements and value sets using the *EHR4CR Terminology Services*. The defined set of eligibility criteria are represented as SPARQL queries. For querying heterogeneous CDWs through a standard interface, we adopted the «A_SupportingClinicalStatementUniversal» model, component of the StudyDesign, proposed by the HL7 Regulated Clinical Research Information Model (RCRIM) Work Group [12]. Structures and Value Sets of Common Data Elements (CDEs) are defined in order to specify additional constraints on the high-level EHR4CR information model and to represent the fine-grained clinical information included in eligibility criteria constructs. Based on the pre-defined terminology mappings in the terminology server, we expand and transform the initial SPARQL queries based on the local CDW schemas, and execute them across different CDWs. We transform the query results based on the standardized terminologies, and then display them at the *User Workbench*. The key elements are discussed as follows.

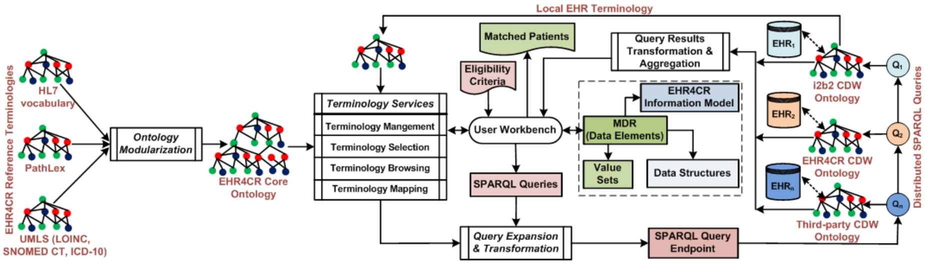

Figure 1. EHR4CR Semantic Interoperability Framework for eligibility determination

3.1 EHR4CR Semantic Resources: Extracting modules from standardized clinical terminologies

Standard clinical terminologies – such as SNOMED CT, LOINC, ICD-10, PathLex, ATC, etc – contain numerous clinical concepts usually organized in a hierarchy and interconnected by domain specific relations, provides a comprehensive medical knowledgebase. However, in particular use-cases and applications only a small fraction of these terminologies are required. Therefore, there have been efforts towards modularizing and extracting disease-specific ontology modules [13]. In addition, Del Vescovo et. al. demonstrated a large scale investigation into the modular structure of state-of-the-art biomedical ontologies, by using their proposed notion of *atomic decomposition*, and extracting relevant and "logically complete" fragments from source ontologies [14]. In order to incorporate the use of standard clinical terminologies in EHR4CR, we plan to investigate the state-of-the art modularization approaches and utilize them for extracting modules that represents those terminology fragments that cover only our domains of interest – i.e. Patient Care and Clinical Research. The extracted terminology modules will be uploaded into EHR4CR terminology server.

3.2 Building SPARQL Query Endpoint based on heterogeneous CDW Models

For building a SPARQL endpoint, Mate et. al. demonstrated the use of SW standards in populating i2b2-based CDWs from heterogeneous EHRs [15]. In EHR4CR, we query patient data from heterogeneous CDWs using SPARQL endpoints that directly expose the CDW schema as RDF. Immediately formalizing the relational database (RDB) model of EHR into a CDW represented as RDF triplestore creates the possibility to use SW technologies – such as defining rules for reasoning and performing *Query Result Transformation and Aggregation*. The RDB-to-RDF mapping [16] ensures the same semantics between the SPARQL endpoints and the RDB, and in this case provides a one-to-one mapping, where: (i) a database table is mapped to an RDFS class (rdfs:Class), (ii) a database table column is mapped to an RDF property (rdf:Property), and (iii) the database data type of a field is mapped to the XSD data type range class of the property; with an exception: if a field is a foreign key, its range is the class that the foreign key points to. Consequently the process of creating an ontology that describes the CDW model can be automated as discussed in [17].

3.3 EHR4CR Semantic Interoperability Services

In order to establish semantic interoperability between data elements describing eligibility criteria and patient data, we are currently developing the following services.

3.3.1. Terminology Services
For constructing the user-defined eligibility criteria at the Workbench, we build terminology services for selecting preferred medical terminologies and value sets:

- *Terminology Management Service* incorporates extracted modules from standard clinical terminologies (LOINC, SNOMED CT, ICD-10, HL7 vocabulary, PathLex, ATC) by loading their schema models through standard protocols.
- *Terminology Selection Service* allows users to select the preferred terminology in which the user wants to define the eligibility criteria.
- *Terminology Browsing Service* allows users to browse the appropriate terminology concepts and attached value sets for defining eligibility criteria.
- *Terminology Mapping Service* manages and provides concept mappings between EHR4CR terminology and local EHR/CDW terminologies.

3.3.2. Query Expansion & Transformation Service

In different clinical terminologies, medical concepts are organized with varied granularities, therefore while querying based on heterogeneous clinical terminologies, query expansion becomes a crucial task [18]. In our work, using the EHR4CR terminology services, we perform query expansion on the user-defined SPARQL queries by walking through terminology hierarchies for a specific terminology concept to incorporate its narrower concepts (i.e. sub-concepts) into the query set. By using the *Query Expansion & Transformation Service*, we transform the expanded SPARQL queries based on local CDWs terminology, which can then be executed across different CDWs to obtain more comprehensive query results.

3.3.3. Result Transformation and Aggregation Service

This service is designed to translate back the query-results obtained from various CDWs into an integrated result format based on the standardized medical vocabulary representing the initially given eligibility criteria. It invokes *Terminology Mapping Services* to retrieve mappings from local to central terminology codes, and also performs necessary unit and measurement conversions among query results.

4 Discussion & Conclusion

In this paper, we present EHR4CR semantic interoperability approach for bridging the clinical care and clinical research domains. In course of developing this framework, we face several challenges: (i) formally defining patient eligibility criteria including temporal constraints [19], (ii) dealing with heterogeneity between different EHRs, (iii) defining mappings between data elements from eligibility criteria and patient data, and (iv) investigating standard query interfaces for retrieving patient information from heterogeneous EHRs. We also need to continue our efforts at harmonizing the EHR4CR Information Model, common data elements and terminology to other standard-based semantic resources including FHIR, BRIDG, CDISC SHARE (and other meta data repository initiatives such as caDSR, openMDR, eMERGE) and the Ontology of Clinical Research (OCRe) [9].

Our preliminary investigations of 11 hospital EHR systems across Europe suggest variable quality in the collection of data items most frequently used as eligibility criteria. However, taken together, queries across multiple eligibility criteria are expected to result in a significant improvement in the accuracy of patient number estimations than present approaches. We will be conducting these and other evaluations more formally later in the project, and will report the results in later publications.

Acknowledgements. The research leading to these results has received support from the Innovative Medicines Initiative Joint Undertaking under *grant agreement* n° [*No 115189*], resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies' in kind contribution.

References

1. Prokosch HU, Ganslandt T.: Perspectives for medical informatics. Reusing the electronic medical record for clinical research. Methods Inf Med. 48(1):38-44 (2009).
2. Fridsma DB., Evans J., Hastak S., Mead CN.: The BRIDG project: a technical report. J Am Med Inform Assoc.15(2):130-7 (2008).
3. El Fadly A., Rance B., Lucas N., Mead C., Chatellier G., Lastic P.Y., Jaulent M.C., Daniel C.: Integrating clinical research with the Healthcare Enterprise: From the RE-USE project to the EHR4CR platform. J Biomed Inform. Dec;44 Suppl 1:S94-S102 (2011).
4. Electronic Healthcare Record for Clinical Research (EHR4CR) [Online]. Available: http://www.ehr4cr.eu/
5. Weber GM, Murphy SN, McMurry AJ, Macfadden D, Nigrin DJ, Churchill S, Kohane IS.: The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories. J Am Med Inform Assoc. 16(5):624-30 (2009).
6. PONTE Project [Online]. Available: http://www.ponte-project.eu/
7. Antoniades A, Georgousopoulos C, Forgo N, Aristodimou A, Tozzi F, Hasapis P, Perakis K, Bouras T, Alexandrou D, Kamateri E, Panopoulou E, Tarabanis K, Pattichis C. Linked2Safety: A secure linked data medical information space for semantically-interconnecting EHRs advancing patients' safety in medical research, International Conference on Bioinformatics and Bioengineering, Larnaka, Cyprus, 2012.
8. EURECA Project [Online]. Available: http://eurecaproject.eu/
9. Sim I., Carini S., Tu S., Wynden R., Pollock B.H., Mollah S.A., Gabriel D., Hagler H.K., Scheuermann R.H., Lehmann H.P., Wittkowski K.M., Nahm M., Bakken S.: The human studies database project: federating human studies design data using the ontology of clinical research. AMIA Summits Transl Sci Proc. Mar 1;2010:51-5 (2010).
10. Fast Health Interoperability Resources (FHIR). [Online]. Available: http://www.hl7.org/implement/standards/fhir/fhir-book.htm
11. W3C task on Clinical Observations Interoperability (COI) [Online]. Available: http://esw.w3.org/HCLS/ClinicalObservationsInteroperability http://www.w3.org/blog/hcls/
12. HL7 Regulated Clinical Research Information Model (RCRIM) Work Group [Online]. http://wiki.hl7.org/index.php?title=Regulated_Clinical_Research_Information_Management
13. Milian, K.; Aleksovski, Z.; Vdovjak, R.; Teije, A. ten ; and Harmelen, F. van 2.: Identifying disease-centric subdomains in very large medical ontologies. Or: finding 2500 out of 300.000. workshop on Knowledge Representation for Healthcare, Springer Verlag (2009).
14. Del Vescovo C, Gessler D, Klinov P, Parsia B, Sattler U, Schneider T, and Winget A. Decomposition and Modular Structure of BioPortal Ontologies. International Semantic Web Conference (1): 130-145 (2011).
15. Mate S, Bürkle T, Köpcke F, Breil B, Wullich B, Dugas M, Prokosch HU, Ganslandt T.: Populating the i2b2 database with heterogeneous EMR data: a semantic network approach. Stud Health Technol Inform. 169:502-6 (2011).
16. R2RML: RDB to RDF Mapping Language. http://www.w3.org/TR/r2rml/
17. Sun H, Depraetere K, De Roo J, De Vloed B, Mels G, Colaert D. Semantic integration and analysis of clinical data. eprint arXiv:1210.4405 (2012).
18. Bettembourg C., Diot C., Burgun A., Dameron O.: GO2PUB: Querying PubMed with semantic expansion of gene ontology terms. Journal of Biomedical Semantics 3:7 (2012).
19. Boland MR., Tu SW, Carini S., Sim I., Weng C.: EliXR-TIME: A Temporal Knowledge Representation for Clinical Research Eligibility Criteria, Proc of AMIA 2012 Clinical Research Informatics Summit (2012).