COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 118 (2015) 322–329

journal homepage: www.intl.elsevierhealth.com/journals/cmpb

Enabling semantic interoperability in multi-centric clinical trials on breast cancer

Raul Alonso-Calvo[a,*], *David Perez-Rey*[a], *Sergio Paraiso-Medina*[a], *Brecht Claerhout*[b], *Philippe Hennebert*[c], *Anca Bucur*[d]

[a] *Biomedical Informatics Group, DLSIIS & DIA, Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegancedo S/N, 28660 Boadilla del Monte, Madrid, Spain*
[b] *Custodix NV, Kortrijksesteenweg 214b3, Sint-Martens-Latem, Belgium*
[c] *Institut Jules Bordet, 121 Boulevard de Waterloo, Brussels, Belgium*
[d] *PHILIPS Research Europe, High Tech Campus 34, Eindhoven, Netherlands*

ARTICLE INFO

Article history:
Received 22 August 2014
Received in revised form
10 December 2014
Accepted 23 January 2015

Keywords:
Semantic interoperability
Data integration
Clinical trials
Clinical research informatics
HL7
SNOMED CT

ABSTRACT

Background and objectives: Post-genomic clinical trials require the participation of multiple institutions, and collecting data from several hospitals, laboratories and research facilities. This paper presents a standard-based solution to provide a uniform access endpoint to patient data involved in current clinical research.
Methods: The proposed approach exploits well-established standards such as HL7 v3 or SPARQL and medical vocabularies such as SNOMED CT, LOINC and HGNC. A novel mechanism to exploit semantic normalization among HL7-based data models and biomedical ontologies has been created by using Semantic Web technologies.
Results: Different types of queries have been used for testing the semantic interoperability solution described in this paper. The execution times obtained in the tests enable the development of end user tools within a framework that requires efficient retrieval of integrated data.
Conclusions: The proposed approach has been successfully tested by applications within the INTEGRATE and EURECA EU projects. These applications have been deployed and tested for: (i) patient screening, (ii) trial recruitment, and (iii) retrospective analysis; exploiting semantically interoperable access to clinical patient data from heterogeneous data sources.

1. Introduction

Current oncology research aims to translate the explosion of genomic research results into clinical practice. Molecular tests are being introduced in modern clinical trials, increasing the potential number of variables and trials [1]. The participation of multiple institutions and semantic interoperability has become therefore essential for the feasibility and sustainability of post-genomic clinical trials.

Although different types of semantic interoperability have been defined in the literature [2], two main topics from previous initiatives in the area has been considered in this work: (i) different data models – such as the HL7 Reference Information

* *Corresponding author*. Tel.: +34 913363748.
E-mail address: ralonso@infomed.dia.fi.upm.es (R. Alonso-Calvo).
http://dx.doi.org/10.1016/j.cmpb.2015.01.003
0169-2607/

Model [3], Informatics for Integrating Biology and the Bedside (i2b2) [4], Observational Medical Outcomes Partnership (OMOP) [5] or CaGRID [6] – and (ii) different vocabulary representations – such as SNOMED CT [7], LOINC [8], HGNC [9], ICD [10], or NCIt [11] – of legacy systems from institutions involved in the same trial. These challenges were identified by analyzing data sources available in the EU projects INTEGRATE: "*Driving excellence in integrative cancer research*" [12] and EURECA: "*Enabling information re-Use by linking clinical Research and CAre*" [13]. In practice, such heterogeneities are solved mainly manually and involve a significant amount of the clinical trial information management costs.

This paper proposes a novel semantic interoperability approach based on interoperability standards and vocabularies [14]. A binding process has been created in order to automatically relate terminologies with data models [15]. While a solution leveraging on the latest semantic technologies and normalization procedures following HL7 and SNOMED CT recommendations has been implemented [16]. The proposed solution is based on a modular service-oriented architecture to provide a homogeneous patient data endpoint. Regarding the architectural approach, we have discarded a solution that completely relies on query translation, to avoid performance issues and to facilitate complex data transformation required in the biomedical domain [2]. Previous initiatives also provide a common data model [4–6] that can be deployed in different instances, but sharing query mechanisms and applications.

The main contributions of this work are: (i) an automatic mechanism to bind biomedical vocabularies and data models, solving vocabulary overlapping issues among others; and, (ii) a normalization mechanism implemented within performance boundaries suitable for multi-centric clinical trial scenarios. The proposed semantic interoperability layer has been tested by application prototypes that homogeneously access clinical data within INTEGRATE and EURECA projects. Concretely, for three main scenarios: (i) patient screening, to find the optimal trial that best fits the needs of a given patient, (ii) trial recruitment, where the aim is to find suitable patients for a given clinical trial. Both patient screening and trial recruitment scenarios share technical use cases that need to match clinical data of patients with eligibility criteria. And finally, (iii) the retrospective analysis of multiple clinical trials requires the definition of patient cohorts for further study, addressing important data sharing challenges [17]. In addition, such scenarios require the exploitation of knowledge stored in biomedical vocabularies, mainly by relationships among concepts. Therefore, based on existing standards, we have introduced Semantic Web technologies to extract such knowledge, exploited afterwards in clinical data retrieval.

2. Methods

The proposed solution aims to homogenize shared information from Clinical Trial Management Systems (CTMS), Laboratory Information Management Systems (LIMS) and Electronic Health Record (EHR) systems among others. A Common Information Model (CIM) is therefore required to provide a semantic interoperability layer across systems. As depicted in Fig. 1, two main components were created within the CIM: (i) the Common Data Model (CDM) and (ii) Core Dataset.

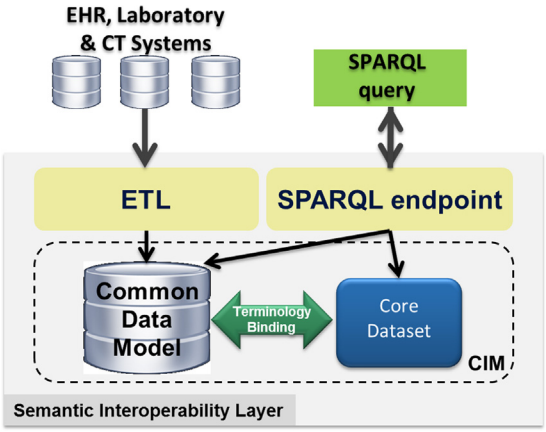

Fig. 1 – Main components of the proposed semantic interoperability layer.

The CDM is the structure and concepts required to homogenize ad-hoc data models of information systems from different institutions – or even in departments of the same organization. Built from subsets of clinical terminologies that fully represent different data sources and scenarios [25], the Core Dataset includes such concepts relationships among them. Finally, the Terminology binding is necessary to define relations between the CDM and the Core Dataset.

As it is depicted in Fig. 1, CDM is populated from heterogeneous databases using Extract, Transform and Load (ETL) processes. These transformations processes are explained elsewhere [14].

2.1. *Common data model*

The main requirement for the CDM is that any data source from clinical institutions must be able to be efficiently represented – in performance and usability terms. The HL7 version 3 Reference Information Model (RIM) has been selected as the foundation of the CDM. The HL7 RIM offers a wide coverage for representing clinical data and has proven useful for clinical information exchange [3]. Moreover, built-in tools from legacy systems, or open source HL7 mapping tools – such as Mirth Connect [18] – for generating HL7 v3 messages can be used.

Although the RIM is the core element of the HL7 v3 standard, it is not a definition of a database structure but an object model. Due to the coverage of the RIM – that also includes specification for messaging, financial data, imaging data, etc., only some RIM classes have been implemented to represent patient clinical data of breast cancer clinical trials:

- Act, with the sub-classes *Observation*, *Procedure*, *Substance Administration*, and *Exposure*.
- Role
- *Entity*, with the sub-classes *Living Subject*, *Person*, and *Device*

Two main relationships classes from the RIM were also required: