Public Health and Informatics
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210116

Towards a Semantic Data Harmonization Federated Infrastructure

Catalina MARTINEZ-COSTA[a,b,1] and Francisco ABAD-NAVARRO[a,b]
[a] *University of Murcia, Murcia, Spain*
[b] *Biomedical Research Institute of Murcia (IMIB-Arrixaca), Murcia, Spain*

Abstract. Data integration is an increasing need in medical informatics projects like the EU Precise4Q project, in which multidisciplinary semantically and syntactically heterogeneous data across several institutions needs to be integrated. Besides, data sharing agreements often allow a virtual data integration only, because data cannot leave the source repository. We propose a data harmonization infrastructure in which data is virtually integrated by sharing a semantically rich common data representation that allows their homogeneous querying. This common data model integrates content from well-known biomedical ontologies like SNOMED CT by using the BTL2 upper level ontology, and is imported into a graph database. We successfully integrated three datasets and made some test queries showing the feasibility of the approach.

Keywords. Ontologies, SNOMED CT, semantic harmonization, graph database

1. Introduction

Despite ethical and technological challenges, clinical data sharing is pursued with increasing efforts to improve research and clinical outcomes. Here we focus on the technological challenges regarding the syntactic and semantic integration of data from multidisciplinary sources. Such data is usually heterogeneously structured (from free text to structured data) and might use different languages. In addition, meaning discrepancies might exist across institutions, which requires syntactic and semantic harmonization.

This is the case of the EU Precise4Q project [1] on stroke management, for which we are implementing the data harmonization infrastructure. Most existing clinical data architectures load data into a common integrated repository within an institution [2]. However, in collaborative, cross-institutional projects as in Precise4Q, institutions' data sharing agreements do not allow data to leave their source repositories. This requires federated architectures in which several repositories are integrated virtually by sharing a common data representation layer, which allows their homogeneous querying. The proposed data harmonization architecture has three main components: (1) semantic common data model (CDM); (2) graph-based repository and a (3) semantic query system. This work focuses on describing (1) and (2).

We depart from anonymized and quality assured data extracted from each source repository. This data is then transformed into a semantically rich representation and harmonized. While multiple data modelling standards and proposals exist to integrate

[1] Corresponding Author, Catalina Martínez-Costa, E-mail: cmartinezcosta@um.es.

and share data, they still fail in providing semantically rich representations [3]. The semantic CDM proposed integrates content from well-known biomedical ontologies like SNOMED CT (SCT) by using a biomedical top-level ontology. The CDM is agnostic to existing clinical modelling representations (e.g. i2b2, OMOP CDM, HL7 FHIR, etc.) and can be extended to represent specific data modelling use cases in a standardized way.

The model consists of a collection of OWL ontologies [4], which are imported into a graph database as a labelled property graph (LPG) for performance reasons.

In the following we describe the main components of the semantic data harmonization infrastructure and the modelling and integration of some data samples into the semantic repository.

2. Materials and Methods

2.1. Semantic Common data model

The integration of heterogeneous data requires of a common data model (CDM) to provide a uniform and unambiguous representation. Semantic integration is the highest integration level and aims to preserve "the detail, uncertainty, and above all the context of the data involved" [5]. Nowadays the vast majority of CDMs are not able to provide this level of data integration. Examples of CDMs for clinical data are openEHR RM [6], HL7 FHIR resources [7] or the OMOP CDM [8] from the OHDSI community. In the past years, efforts have been made to improve their semantic capabilities.

Current semantic integration approaches are based on ontologies. As mentioned in [3], well-designed ontologies, which focus on logical consistency and represent concepts within a hierarchy are useful for the creation of clinical data representations that are semantically rich, unambiguous, and not dependent on specific CDMs.

The proposed ontology infrastructure follows the semantic harmonization principles from Cunningham et al. [9]. It uses the axiomatically rich top-level ontology BTL2 [10] as reference harmonization framework to allow the unambiguous integration of domain-specific knowledge. The CDM is agnostic to existing clinical data modelling specifications and includes both information (e.g. temporal context, provenance, etc.) and clinical domain concepts (e.g. "stroke", "obesity", etc.). SNOMED CT [11] is our ontology of choice for clinical domain modelling, given its wide coverage of the biomedical domain and its formalization degree that allows logical inference. For each data modelling use case, a local semantic data model is built, with the domain knowledge and by following CDM constraints. Figure 1 describes the ontology infrastructure.

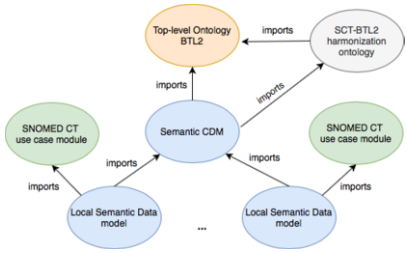

Figure 1. Ontology harmonization infrastructure. In the middle the semantic CDM that uses BTL2 to integrate clinical and information entities [12]. In grey, the partial harmonization of SCT main concepts and attributes with BTL2 [13]. In green SCT modules with use case concepts. At the bottom in blue, local semantic CDMs which comply with the semantic CDM restrictions and include use case domain knowledge.

The central entity of the CDM is the clinical statement that represents documented medical facts (observation, judgment or assessment about a certain patient aspect) as result of a procedure. Other entities such the subject, the provider of information or the healthcare process provide context to the documented medical fact (see Figure 2).

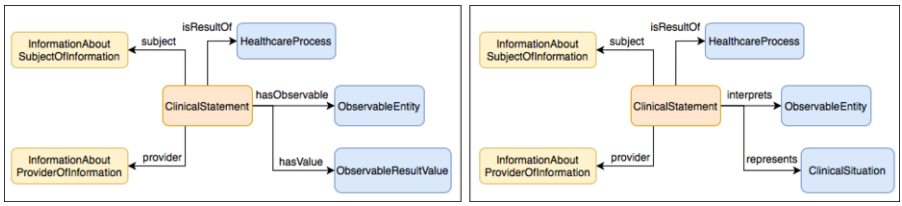

Figure 2. Excerpt of two schemas for clinical statement modelling. Left: a medical fact is described by two concepts describing what is observed and its result respectively. Right: a medical fact is described by one concept which describes the result of the observation / assessment. Classes in blue represent medical domain concepts (e.g. obese clinical finding in SCT as *ClinicalSituation*, weight and body mass assessment procedure in SCT as *HealthcareProcess*). Information entity classes in yellow.

2.2. Graph-based database

Graph databases have gained popularity given their performance with complex and interconnected large amounts of data. There are two main graph models, the W3C standard RDF [14] and Labelled Property Graph (LPG). The RDF model uses <subject,predicate,object> triples while the LPG model consists of nodes, relationships and properties. The RDF model provides some advantages over LPGs as data reasoning. However, its adoption outside the academic field is limited due to performance reasons [15]. We have used the Neo4j database [16], which implements the LPG model and is one of the most used graph databases [17]. In order to get RDF-like advantages and performance from Neo4j:

1) For each dataset, a local semantic data model is built by following CDM constraints
2) Mappings between data and local semantic data model entities are defined and RDF instances are created accordingly
3) Data reasoning infers new statements. Transitivity along concept hierarchies is applied to make for every resource, *rdf:type* together with its superclasses explicit.
4) Ontology population into Neo4j by translating OWL/RDF into LPG with Neosemantics [18]. It imports a simplified ontology version in the graph and enables queries with taxonomic reasoning.
5) Instance data population into Neo4j with [18] and by following RDF to LPG translation rules [19].

3. Results

The semantic harmonization process (modelling and integration) uses data excerpts from stroke patients at a rehabilitation institute provided as anonymized CSV files.

3.1. Data modelling

For representing data according to the semantic CDM. a local semantic data model is built for each dataset with the specific domain knowledge, supported by domain experts

and data owners, and following the CDM constraints. Whenever possible, concepts from SCT are used for representing domain knowledge. If no concept exists neither in SCT nor in any other clinical vocabulary (e.g. LOINC, ICD, etc.) a new one is crated as a child concept of an existing one of the local semantic data model. E.g. at one dataset, results from psychometric orientation tests are classified in three types: personal, spatial and temporal orientation. SCT provides only *405017007|Ability to identify person, place, and time (observable entity)|*. Thus, three subconcepts are created and the SCT orientation concept inherits from each one.

There are some complex modelling cases, such as the representation of the patients' age of stroke onset, which could only represented by complex OWL definitions. Aware of possible limitations but in favor of performance and functionality, we decided to encapsulate the entire meaning in a concept similarly as SCT does with concepts like *170092006 | Age when first sat (observable entity)|*. The new concept is placed under the closest standardized one in the ontology. Figure 3 shows an excerpt of the modelling of the result of the Stroop word test according to semantic CDM.

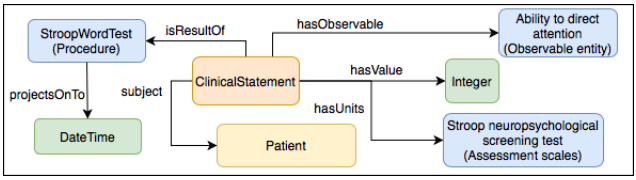

Figure 3. Excerpt local semantic data model for representing the result of the cognitive "Stroop word test" according to CDM (see Fig.3 (left)). In blue SCT concepts. In green, literal values to be instantiated

3.2. Data integration

Once local semantic models are built for each dataset and steps 1) to 5) performed (see section 2.2), they are instantiated and loaded into the graph database. In case datasets have similar data heterogeneously represented, transformation rules informed by domain experts and data owners are implemented and the corresponding data instances created. E.g. body mass index in one dataset vs. obesity = (yes/no) in another one.

When having distinct datasets about same patients, one of the benefits of the graph-based representation is that all data is represented in one graph and thus, queries including aspects from them can be executed in a more efficient and intuitive way as in a classical relational database. In addition, taxonomic reasoning allows queries to be performed at different granularity levels: E.g. retrieve for all patients whose orientation ability was initially evaluated, the number of orientation tasks they performed in the rehabilitation therapy. This query involves two different datasets and taxonomic reasoning for getting all orientation assessments (see section 3.1).

4. Discussion and Conclusion

A semantic data harmonization infrastructure that can be implemented as a federated architecture has been proposed. It consists of (1) a semantic CDM, (2) a graph-based database and a (3) semantic query system. This work focuses on (1) and (2). For each dataset at a source site, a local semantic data model is built supported by the data experts and following CDM constraints. The semantic CDM includes a collection of

standardized ontologies under a top-level ontology for harmonization. This CDM is agnostic regarding existing standardized models, focusing on data meaning representation and acting as their common model to allow data transformations among them. At each data site the local semantic model is instantiated and loaded into a graph database according to the steps in section 2.2. A direct benefit of this semantic data enrichment and its representation as a graph is the provision of a harmonized data representation across the heterogeneous data sites. In addition, queries including aspects from several datasets can be executed in a more efficient and intuitive way as in classical relational databases and exploit taxonomic reasoning. Neo4j with Cypher as query language have proven useful. We are currently implementing a semantic query system based on the proposed semantic CDM, customizable for each data integration scenario. We have already harmonized and integrated three datasets into the graph database, what shows the feasibility of the approach. The highest workload is the building of each local semantic data model and the mapping definition, which requires both knowledge on the data and the target data model. Nevertheless, once implemented it can be reused to import new data and on the other hand, such tasks need support of domain experts and cannot be fully automatized.

Acknowledgments: Precise4Q (GA: 777107) H2020-SC1-2017-CNECT-2; DATA4HEALTH. RTI2018-099039-J-I00 (Retos 2018)

References

[1] Precise4Q project, Available at: https://precise4q.eu , Accessed Jan., 2021.
[2] Gagalova KK, Elizalde MA, Portales-Casamar E, Görges M. What You Need to Know Before Implementing a Clinical Research Data Warehouse. JMIR formative research. 2020 Aug 27; 4(8):e17687
[3] Freedman HG, Williams H, Miller MA, Birtwell D, Mowery DL, Stoeckert Jr CJ. A novel tool for standardizing clinical data in a semantically rich model. J. Biomed. Informatics: X. 2020 Dec 1;8:100086.
[4] Semantic Common Data Model. Available at: https://github.com/P4Q-UM/SemanticDataModel Accessed Jan, 2021.
[5] Cheatham M, Pesquita C. Semantic Data Integration. In: Zomaya AY, Sakr S, editors. Handbook Big Data Technologies. Cham: Springer International Publishing; 2017. p. 263–305.
[6] OpenEHR RM, Available at: https://specifications.openehr.org/releases/RM/latest, Accesed Jan, 2021.
[7] HL7 FHIR resources, Available at: https://www.hl7.org/fhir/resourcelist.html. Accessed Jan., 2021.
[8] OMOP CDM, https://www.ohdsi.org/data-standardization/the-common-data-model/,Accessed Jan. 2021.
[9] Cunningham JA, Van Speybroeck M, Kalra D, Verbeeck R. Nine principles of semantic harmonization. In AMIA Annual Symposium Proceedings: AMIA; 2016. p. 451.
[10] Schulz S, Boeker M, Martinez-Costa C. The BioTop family of upper level ontological resources for biomedicine. Stud Health Technol Inform. 2017 Jan 1;235:441-5.
[11] SNOMED International. Available at: https://www.snomed.org, Accessed Jan. 2021.
[12] Schulz S, Martínez-Costa C, Karlsson D, Cornet R, Brochhausen M, Rector AL. An Ontological Analysis of Reference in Health Record Statements. InFOIS 2014 Sep 5 .p. 289-302.
[13] Schulz S, Martínez-Costa C. Harmonizing SNOMED CT with BioTopLite: An Exercise in Principled Ontology Alignment. InMedInfo 2015 .p. 832-836.
[14] Resource Description Framework (RDF). Available at https://www.w3.org/RDF/, Accessed Jan. 2021.
[15] Angles R, Prat-Pérez A, Dominguez-Sal D and Larriba-Pey J.L. Benchmarking database systems for social network applications. In Grades 2013 international workshop. 2013.p. 1-7).
[16] Neo4j, Available at https://neo4j.com, Accessed Jan. 2021.
[17] Guia J, Soares VG, Bernardino J. Graph Databases: Neo4j Analysis. InICEIS (1) 2017 Jan. p. 351-356.
[18] Neosemantics, Available at https://neo4j.com/labs/neosemantics/, Accessed Jan. 2021.
[19] Bouhali R, Laurent A. Exploiting RDF open data using NoSQL graph databases. InIFIP International Conference on Artificial Intelligence Applications and Innovations; Springer, Cham.2015 Sep 14.p. 177-190.