EMIF Deliverable 11.4: First complete version of the harmonised information model and associated terminology mapping

Executive summary

Executive Summary

This is the fourth deliverable of work package 11, which focuses on the support of interoperability across the EMIF federation ecosystem, in particular focusing on semantic interoperability. The challenge being addressed by this work package is to enable research queries that will be expressed in terms and concepts familiar with the research community, to be executed against data that originates from diverse data sources. These data sources will each have potentially quite different data structures, but more importantly will have different defined data items with different choices of terminology, measurement units, etc.

The previous deliverables from this work package have summarised the state-of-the-art approaches being pursued to this challenge by both the eHealth and clinical research communities globally. The needs of this project have been explained, and early experimentation in methods to tackle the challenge have been reported.

In parallel to the ongoing work of this work package, the project as a whole has progressively crystallised more and more clearly the nature of the research services that the EMIF platform will aim to provide, and therefore the kinds of research queries and the extent of harmonisation of the the data that needs to underpin their execution. An important decision that was made last year across multiple platform work packages has been to prioritise the implementation of suitability and feasibility queries, and to adopt the OHDSI / OMOP Data Model as the "common data model" of the project. Irrespective of whether the underlying data originates from routinely collected electronic health records, or from dedicated data collection on a population cohort, there is a need for a harmonised common view of the key data items, and summary statistics about the values held about each one, to enable suitability and feasibility queries. This is the role for which OMOP has been selected. This decision has been the starting point for the work reported in this deliverable, and the first section of the deliverable summarises this model.

As a consequence of this decision, the challenge of semantic harmonisation becomes focused on the ETL process, whereby heterogeneous data are mapped to a harmonised schema that can be used as the basis for import into the common data model. This is the first priority for semantic harmonisation, but the approaches being developed here will also be directed towards the harmonisation of fine-grained datasets for analysis (which might not always use OMOP).

As previous work package 11 deliverables have described, the EMIF Knowledge Object is the semantically formalised definition of a discrete data item or cluster of data items that represent a single health or health related concept, which may be defined from a clinical point of view, life sciences point of view or from a research point of view. The knowledge object is the standardised representation of some clinical data item or data item cluster, which may either faithfully represent the corresponding data within one

This work has received support from the EU/EFPIA Innovative Medicines Initiative Joint Undertaking (EMIF grant n° 115372).

data source (a local knowledge object), or may represent the harmonised view of a set of equivalent local knowledge objects, a global knowledge object, which will be the virtual representation used to fashion research queries and to configure result sets. The transformations needed to go between local knowledge objects and global knowledge objects are mapping rules that must also be formalised and capable of computable execution.

Through working closely with the AD vertical in the project a number of very important guiding principles have been identified for the way in which knowledge objects should be constructed and used. Section 2.1 presents these guiding principles.

Practical implementation work has been undertaken over the past year on the formalised representation of knowledge objects (using Description Logic), and the tools needed to compose them, manipulate them, to add access control and provenance information to the basic semantic representations (using Topbraid Composer), and to formalise the representation of transformation rules (using SPIN). Section 2.2 presents the high-level structure of the ontology containing the knowledge objects. Section 2.3 summarises the main classes defining the knowledge object library, and Section 2.4 documents how the mapping rules are defined. Sections 2.5 and 2.6 describe how this semantic framework can be applied at a local level by each data source. The last sections of the document describe the practical work undertaken in establishing the initial knowledge object library with AD example content.

From definitions that required intensive work by informatics semantic specialists, the work of the last year has shifted the balance of effort closer to the domain experts, which is appropriate. Further work to be undertaken next year will focus on making the tools presented to the domain experts more friendly and minimising their required effort to define the knowledge objects they need to undertake their areas of research. Further work will also be undertaken to optimise the definition and execution of mapping rules, and to facilitate the work of local data managers in generating and mapping local knowledge objects. The focus, therefore, will increasingly be on scaling up the usability of this approach.

Contacts

EMIF-Platform: Johan van der Lei – Nigel Hughes
j.vanderlei@erasmusmc.nl - nhughes@its.jnj.com

This work has received support from the EU/EFPIA Innovative Medicines Initiative Joint Undertaking (EMIF grant n° 115372).