A Framework for Evaluating and Utilizing Medical Terminology Mappings

Sajjad HUSSAIN, Hong SUN, Anil SINACI, Gokce Banu Laleci ERTURKMEN, Charles MEAD, Alasdair J. G. GRAY, Deborah L. McGUINNESS, Eric PRUD’HOMMEAUX, Christel DANIEL, Kerstin FORSBERG

Abstract. Use of medical terminologies and mappings across them are considered to be crucial pre-requisites for achieving interoperable eHealth applications. Built upon the outcomes of several research projects, we introduce a framework for evaluating and utilizing terminology mappings that offers a platform for i) performing various mappings strategies, ii) representing terminology mappings together with their provenance information, and iii) enabling terminology reasoning for inferring both new and erroneous mappings. We present the results of the introduced framework from SALUS project where we evaluated the quality of both existing and inferred terminology mappings among standard terminologies.

Keywords. Terminology mapping, semantic interoperability, reasoning, validation

Introduction

Achieving a computable semantic interoperability (CSI) among different healthcare applications is – at its core – deeply dependent on the use of “controlled terminologies” which enable the inter-machine exchange of clear and computationally unambiguous semantics [1]. Aiming towards this goal, clinical experts go through a process of defining “terminology mappings” between “standard” terminologies developed by standards organizations (CDISC, IHTSDO, ICH, etc.), as well as local/legacy terminologies to support a number of specific CSI use cases including (but not limited to): (i) semantic interoperability between clinical care and clinical research systems for pharmacovigilance and patient safety, (ii) clinical decision making, and (iii) clinical data integration and mediation in pharmaceutical R&D.

Recent research projects focusing on CSI—such as SALUS [2], Open PHACTS [3], and EHR4CR [4]—have published their experiences in defining and utilizing multiple mappings between various terminologies [5]. Although, on the surface, it may
appear to the uninitiated as a simple exercise like “this term in this terminology is the same as that term in that terminology”. However, it is often a considerably challenging task [1] due to: (i) availability of up-to-date information to assess the suitability of a given terminology for a particular use case; (ii) difficulty of correctly using complex, rapidly evolving terminologies; (iii) differences in granularity between the source and target terminologies; (iv) lack of semantic mappings in order to completely and unambiguously define computationally equivalent semantics; (v) lack of provenance information, i.e. how, when and for what purposes the mappings were created; and (vi) time and effort required to complete and evaluate mappings. For example, considering SNOMED-CT as hub terminology, both ICD-9-CM and MedDRA codes are mapped to SNOMED-CT codes (see Figure 1). As a result, new mappings are inferred (e.g., “Anaphylactic shock due to serum, not elsewhere classified” (ICD9) ↔ “Anaphylactic transfusion reaction” (MedDRA)). However, via reasoning, problematic or incorrect mappings are also found (see Figure 1: highlighted in red).

Another important problem is the inclusion of implicit context in the use of a particular term which can result in certain usage contexts with terms that are semantically equivalent in one usage context but not in another [6]. For example, if a generic code for a TNM grade is used in local templates of anatomic pathology reports for cancer, knowing the context—breast or prostate cancer for example—the local code for pT shall be “bound” to the appropriate cancer specific pT code (e.g. either LOINC:44663-3 T classification in Breast tumor or LOINC:44664-1 T classification in Prostate tumor). Taking the context information into account, pT3 value for example has a different value in the context of a breast cancer (Tumor >50 mm in greatest dimension) than in the context of prostate cancer (Extraprostatic extension).

1. Terminology Mappings Evaluation and Utilization Framework

In this paper, we present a framework for evaluating and utilizing medical terminology mappings (see Figure 2). This framework features: (i) exploitation of available terminology mapping and alignment strategies [1, 7] for finding similarities between source/reference terminologies; (ii) presenting the resulted mappings to terminology experts for further validation; (iii) representing the supporting arguments and methods used (i.e. provenance) [8] for defining or finding mappings with a focus on the reason for defining the equivalence relation [9]; (iv) exploiting reasoners to perform
terminology reasoning for finding useful inferred mappings and also to validate both asserted and inferred mappings based on formalized validation schemes and policies [10,11]; and (v) offering RESTful Web Services, REST APIs and query endpoints for querying and utilizing available terminology mappings—based on their provenance—for different contexts or usages in mind.

Figure 2. Terminology Mappings Evaluation and Utilization Framework.

1.1. Terminology Mapping Strategies and Mapping Provenance Representations

There have been various efforts in finding and representing mappings between standard medical terminologies [1,7]. BioPortal, a repository of biomedical terminologies/ontologies, with more than 300 ontologies to date, performs a variety of mapping algorithms and strategies [7] include: (i) Lexical Mappings (LOOM), represented via skos:closeMatch property, generated by performing lexical comparison between preferred labels and alternative labels of terms; (ii) Xref OBO Mappings, represented via skos:relatedMatch property, processed by using Xref and Dbxref are properties used by ontology developers to refer to an analogous term in another vocabulary; (iii) CUI Mappings from UMLS, represented via skos:closeMatch property, are extracted by utilizing the same Concept Unique Identifier (CUI) annotation as join point of similar terms from different vocabularies; (iv) URI-based Mappings, represented via skos:exactMatch property, are generated identity mappings between term concepts in different ontologies that are represented by the same URI.

Existing approaches for exchanging mappings, such as a nanopublication [12] or a VoID linkset [13], focus on capturing and modeling the provenance of the mapping, i.e. what has been mapped, by whom, using what tool. What it does not convey is the context in which the assertion was made, i.e. the operational equivalence that is encoded by the mapping. The Open PHACTS project has extended the linkset approach to include a mapping justification [9]. The linkset header is extended with an additional property that conveys why the mapping holds, e.g. the terms have the same preferred label or that they share an identifier. A key advantage of this approach is that existing reasoning mechanisms, e.g. rules for skos:exactMatch, can be applied without change since the links themselves are represented using a standard ontology. In addition, existing mappings can be simply extended with additional metadata then reused; there is no need to regenerate the linking data [14].

1.2. Terminology Reasoning and Mapping Evaluation

Once the existing mappings are formally represented, e.g. using SKOS mapping properties [15], terminology reasoning can be applied to derive new mappings and also
to detect both asserted or inferred problematic mappings [11]. Figure 1 shows the inferred mappings between ICD-9-CM and MedDRA by utilizing the existing ICD-9-CM-to-SNOMEDCT and SNOMEDCT-to-MedDRA mappings. When the mappings are large, a statistically sound sample is extracted, and validations are applied on the sample only [1, 10]. It is also important to point out that the terminology reasoning could also lead to conclusions that the mapping creators do not intend to. It is therefore required to perform automated mapping validation schemes to be applied on the entire mappings to detect and present problematic mappings to experts for further investigations and revisions. When the mapping relations and semantic relations can be considered as transitive, it is possible to infer a relation between two concepts in a same terminology system. If such an inferred relation is not stated in (or cannot be inferred from) a terminology system, then inferring such a relation via mapping relations is considered as “vocabulary hijacking” [11]. For example, Figure 1 shows that a problematic mapping (Anaphylactic reaction (MedDRA) ↔ Anaphylactic shock (MedDRA)) is inferred because both concepts are mapped to the same SNOMED-CT concept (39579001). Mapping validations are thus required to detect such mappings that may lead to unintended inferences.

1.3. Terminology Mapping Querying and Utilization

Once terminology mappings are represented together with provenance information, utilizing them to interoperating applications is another challenge. SALUS Terminology Server addresses this challenge and provides a RESTful interface for querying term mappings [16]. SALUS Terminology Server is built upon a triple-store and inherently supports RDF. Term representations are utilized mainly by using SKOS [15]. REST methods returns query results according to this semantic modeling of the terms.

<table>
<thead>
<tr>
<th>Mapping</th>
<th>Source</th>
<th>Target</th>
<th>Hierarchy Relations*</th>
<th>Defined Mappings</th>
<th>Inferred Mappings</th>
<th>Problematic Mappings</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SNOMED-CT</td>
<td>ICD-10</td>
<td>1,839,401</td>
<td>16,838</td>
<td>104,761</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>MedDRA</td>
<td>SNOMED-CT</td>
<td>1,886,575</td>
<td>10,648</td>
<td>453,615</td>
<td>1,790</td>
</tr>
<tr>
<td>3</td>
<td>MedDRA</td>
<td>ICD-10</td>
<td>121,032</td>
<td>139,216</td>
<td>3,331</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>SNOMED-CT</td>
<td>ICD-9-CM</td>
<td>1,845,688</td>
<td>31,862</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>MedDRA</td>
<td>ICD-9-CM</td>
<td>174,493</td>
<td>270,094</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

2. Results

Table 1 presents the defined, inferred and problematic mappings found in the SALUS Terminology Server [5, 16]. Mappings 1, 2 and 4 were created based on the mappings developed by existing projects. These mappings were represented via skos:exactMatch, or exactOrNarrow relations. It can be observed that there are many problematic mappings inferred and detected from already defined SNOMED-CT to ICD-10 mappings (see row 1 in Table 1). It is because of wrong assumptions were made in interpreting source mappings stored in Excel sheet to RDF, therefore inappropriate mapping properties were assigned. Based on the already defined mappings and hierarchy structures of related terminologies, new mappings were
inferred. By using SNOMED-CT as a bridge terminology, the MedDRA-to-ICD10 mappings were derived from the mapping 1 and 2, and the MedDRA-to-ICD9CM mappings were derived from mapping 2 and 4.

3. Conclusion

The role of medical terminologies is vital to achieve CSI in clinical informatics. In this paper, we show the challenging nature of mapping utilization among different terminologies. The introduced framework has been built upon existing terminology mappings to (i) infer new mappings for different CSI use cases, (ii) present provenance of the mappings together with the context information—an important problem for term mapping utilization, and (iii) perform mapping validation in order to show that inferred mappings can be erroneous. The framework enables a more collaborative semantic landscape with providers and consumers of terminology mappings. It can also be the basis for additional services, such as usage data and feedback mechanisms for the providers of mappings and of the source and targets terminologies. Making the use, and reuse, of terminology mappings visible could enable new funding and business models in a sustainable semantic landscape across clinical care and clinical research domains.

References