A web-based tool for development of a Common Ontology between ICD11 and SNOMED-CT

Vincenzo Della Mea, Omar Vuattolo
University of Udine
Udine, Italy
dellamea@uniud.it, omar.vuattolo@uniud.it

Bedirhan T. Ustun
World Health Organization
Geneva, Switzerland
ustunt@who.int

Alan Rector
University of Manchester
Manchester, UK
rector@cs.man.ac.uk

Stefan Schulz
Medical University of Graz
Graz, Austria
stefan.schulz@medunigraz.at

Jean-Marie Rodrigues
Jean Monnet University
Saint Etienne, France
rodrigue@univ-st-etienne.fr

Kent Spackman, David Robinson, Jane Millar
International Health Terminology Standard Development Organisation
Copenhagen, Denmark
ksp@ihtsdo.org, dro@ihtsdo.org, jmi@ihtsdo.org

James R. Campbell
University of Nebraska Medical Center
Omaha, Nebraska, USA
campbell@unmc.edu

Christopher G. Chute, Harold R. Solbrig
Mayo Clinic
Rochester, Minnesota, USA
chute@mayo.edu, solbrig.harold@mayo.edu

Kristina Brand Persson
National Board of Health and Welfare
Stockholm, Sweden
Kristina.Brand.Persson@socialstyrelsen.se

Abstract—Semantic Interoperability, i.e., preserving the meaning among health related data, is one of the crucial topics of Health Informatics. The International Classification of Diseases by WHO and SNOMED-CT, by IHTSDO, are the most prominent systems currently available for coding health data. In 2010 a collaboration agreement between the maintainers of ICD and SNOMED-CT has been signed, and gave birth to a Joint Advisory Group (JAG), aimed at defining a common basis between both terminological systems, in the form of a Common Ontology. In fact, in addition to the base task of defining the logical structure of the Common Ontology, JAG also designed a workflow for identifying its content items. As a preliminary step, the investigation of three specific ICD11 chapters has been designed to understand the total workload and calendar of activities. The present paper briefly describes the method and the tool that has been developed to support the latter investigation.

Keywords—ICD, SNOMED-CT, Ontology, Terminology, Classification, Mapping

I. INTRODUCTION

Semantic Interoperability, i.e., preserving the meaning among health related data, is one of the crucial topics of Health Informatics. The International Classification of Diseases by WHO [1] and SNOMED-CT [2], by IHTSDO, are the most prominent systems currently available for coding health data. To foster semantic interoperability, a collaboration agreement between the maintainers of ICD and SNOMED-CT has been signed in 2010, which eventually drove the birth of a Joint Advisory Group (JAG). Its work is aimed at identifying a common basis between both terminological systems. This culminated in the decision to create a Common Ontology (CO), whose features have been described elsewhere [3]. We here report on a tool that has been developed in an ad-hoc manner to support the initial steps towards the creation of such a Common Ontology, involving in particular the forthcoming version 11 of ICD, which was designed with ontological principles in mind [4] and developed using semantically aware tools, in particular the iCat derivative of WebProtege [5].
In fact, in addition to the base task of defining the logical structure of the Common Ontology, JAG designed a workflow for identifying its content items. As a preliminary step, an investigation on the Circulatory System Disorders chapter of ICD11 helped to set the agenda and identify workflow and possible issues [6]. The experimentation has been then extended to other three specific ICD11 chapters in order to better understand the total workload and calendar of activities.

The present paper briefly describes the method and tool that has been developed to support the latter investigation.

II. THE SYSTEM

A. Requirements

The envisaged tool, called MAPPET, supports an expert terminologist in (i) identifying equivalence relationships between ICD11 representational units (RU) and SNOMED-CT RU and (ii) examining semantic issues that become apparent after having set those equivalences. In fact, while both ICD11 and a subset of SNOMED-CT describe diseases, syndromes, disorders, signs, symptoms and other healthcare-relevant features, they do not always define and organize them in the same way [7].

In addition to equivalences, the tool collects also other kinds of information that could be useful for the development of the Common Ontology:

- ICD11 RUs that do not have an equivalent in SNOMED-CT, to be added to it;
- ICD11 RUs that should not be part of the Common Ontology: residual codes (still in the Foundation Layer, due to the ongoing work on it) or disease groupings;
- ICD11 RUs that need some definition clarification.

The MAPPET tool also maintains record of RUs moved or deleted from ICD11, since ICD11 is still being developed and is under continuous modification.

The typical workflow is as follows:

- The expert terminologist in charge of working with a specific ICD11 chapter downloads from the system the list of chapter RUs, enclosed in an Excel file preformatted with all the fields needed for mapping;
- The file also contains the results of some lexical matching (when possible) that could act as an help for terminologists;
- The terminologists search for equivalences using a SNOMED-CT browser, and report found equivalences in the Excel file by indicating the concept identifier and Fully Specified Name;
- They also report about the above mentioned information for suitable entities;
- When finished, or at any time, they can send back the Excel file to the system maintainers;
- System maintainers upload the provided Excel file on the system database.

![Mappet](image)

MITEL_ University of Udine, Italy for WHO-IHSITDO Joint Advisory Group

<table>
<thead>
<tr>
<th>ICD11</th>
<th>SNOMED-CT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Heart valve disease</strong></td>
<td><strong>Heart valve disease (disorder)</strong></td>
</tr>
<tr>
<td>1. Aortic valve disease</td>
<td>1. Congenital anomaly of heart valve (disorder)</td>
</tr>
<tr>
<td>2. Mitral valve disease</td>
<td>2. Aortic valve disorder (disorder)</td>
</tr>
<tr>
<td>3. Tricuspid valve disease</td>
<td>3. Mitral valve disorder (disorder)</td>
</tr>
<tr>
<td>5. Multiple valve disease</td>
<td>5. Mitral valve regurgitation (disorder)</td>
</tr>
<tr>
<td>6. Aortic valve disease</td>
<td>6. Aortic valve disease (disorder)</td>
</tr>
<tr>
<td>7. Aortic valve stenosis (disorder)</td>
<td>7. Chronic aortic valve disease (disorder)</td>
</tr>
<tr>
<td>8. Aortic valve regurgitation (disorder)</td>
<td>8. Aortic valve regurgitation (disorder)</td>
</tr>
<tr>
<td>9. Aortic valve insufficiency (disorder)</td>
<td>9. Aortic valve insufficiency (disorder)</td>
</tr>
<tr>
<td>10. Aortic valve stenosis (disorder)</td>
<td>10. Aortic valve stenosis (disorder)</td>
</tr>
<tr>
<td>11. Aortic valve regurgitation (disorder)</td>
<td>11. Aortic valve regurgitation (disorder)</td>
</tr>
<tr>
<td>12. Multiple valve disease (disorder)</td>
<td>12. Multiple valve disease (disorder)</td>
</tr>
<tr>
<td>13. Pulmonary valve disease (disorder)</td>
<td>13. Pulmonary valve disease (disorder)</td>
</tr>
<tr>
<td>15. Tricuspid valve disease (disorder)</td>
<td>15. Tricuspid valve disease (disorder)</td>
</tr>
<tr>
<td>17. Aortic valve insufficiency (disorder)</td>
<td>17. Aortic valve insufficiency (disorder)</td>
</tr>
<tr>
<td>18. Mitral valve disease (disorder)</td>
<td>18. Mitral valve disease (disorder)</td>
</tr>
<tr>
<td>19. Pulmonary valve disease (disorder)</td>
<td>19. Pulmonary valve disease (disorder)</td>
</tr>
<tr>
<td>20. Tricuspid valve disease (disorder)</td>
<td>20. Tricuspid valve disease (disorder)</td>
</tr>
</tbody>
</table>

Fig. 1. The ICD11 – SNOMED-CT double browser
Next time the file will be downloaded, it will contain also the latest mappings.

Furthermore, equivalences can be viewed in a graphical form through a web accessible graphical interface.

**Architecture and interface**

The architecture of the system is built around a web-accessible database of equivalences that can be fed through either Excel files or a graphical interface, and can provide output in graphical form, Excel again, or Prolog. The web application is able to maintain multiple equivalences, recorded by author, in order to also study agreement in equivalence identification.

The graphical interface allows for a visualization of a double browser (ICD11/SNOMED-CT) plus a graph of RUs possibly connected by equivalences.

The double browser, shown in Fig.1, shows color-coded RUs, and allows to quickly show the equivalent RU by clicking on the selected RU. In particular, labels highlighted in green point to RUs for which an equivalence is available. The same colour coding reflects on the graphical display, which shows the examined pair of RU, their transitive closure to the roots, and their first-level descendants (fig.2).

Whenever an equivalence or lexical mapping is available, an arc is present in the graph. Extra functions include automatic identification of candidate equivalences through lexical methods, which can help experts in focusing on likely equivalent entities.

Although it initially seems a good idea to use a graphical interface to do the mapping, this approach did not fit well with the work model usually adapted by the terminologists involved, and who have their own methods and tools. Excel was, therefore, chosen for input/output because it is well known by the terminologists and integrates well with their usual tools.

The Prolog output has been used for analysing the equivalences, with the aim of discovering structural issues (e.g., inverted ancestor/descendant relationships, synonyms mapped on different entities, etc.).

The whole system is implemented using PHP, HTML5, Javascript and MySQL on a Linux Ubuntu web server. The graphical display is built using GraphViz and embedded in the web page as SVG.

The ICD11 chapters involved are obtained from the ICD URI API [8], while the corresponding SNOMED-CT subtrees are obtained by importing the international release of January 2014 of the database available from the UMLS.
Terminology Services [9]. While a stable SNOMED-CT is available twice a year, we automatically download the last ICD-11 version once a week, because it is still under development and therefore continuously updated.

<table>
<thead>
<tr>
<th>Table I. Equivalences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Circulatory System</td>
</tr>
<tr>
<td>Equivalences</td>
</tr>
<tr>
<td>No equivalence, to be added to SCT</td>
</tr>
<tr>
<td>No equivalence, not to be added to SCT</td>
</tr>
<tr>
<td>Needs clarification</td>
</tr>
<tr>
<td>ICD11 total entities</td>
</tr>
</tbody>
</table>

III. RESULTS

The system has been applied to the Diseases of Circulatory System chapter in ICD11, on which 3 experts have been involved. The chapter includes a total of 1735 ICD11 entities; the equivalent subtree in SNOMED-CT includes 5562 concepts.

On this first chapter, 681 equivalences have been identified so far, while 209 entities have been identified as unmappable because they are “residuals” (e.g. of the form “other X” or “X not elsewhere classified”) or groupings (e.g. “Endocrine, nutritional and metabolic diseases”). These are appropriate for a classification but not suitable from an ontological point of view. For further details, please check Table 1.

After the first experiment, which also led to the refinement of the software, mapping work started on two other areas of ICD11: the “Ear and mastoid process” chapter, and the “Respiratory system chapter”. Table 1 shows details on the equivalences, and lack of, found until now. Please consider that work is still going on, so most likely these figures will increase.

Using the Prolog description, a few situations have been identified where there is the need for some editorial curation of either ICD11 or SNOMED-CT. As an example, in ICD11 the entity entitled “Congenital anomaly of ear” is descendant of “Structural developmental anomalies of the ear”. The latter in SNOMED-CT can be mapped to “Congenital malformation of ear” that in turn is descendant of “Congenital anomaly of ear”, thus having crossed parenthood (see Fig.3 for a graphical display of the situation).

Finally, agreement among terminologists in charge of determining the equivalences has not yet been studied. However, individual equivalence assignments are maintained in the backend database for further analysis.

![Fig. 3. An example of an issue in knowledge organization discovered through the Prolog version of the equivalences database.](image)
IV. DISCUSSION

There is a clear advantage to harmonising the ICD with SNOMED CT and other ontologically based representations, in order to foster semantic interoperability among health information systems.

The tool discussed in the present paper is currently allowing proceeding forward with the first step towards a Common Ontology between ICD and SNOMED-CT, by individuating those entities that, being equivalent, do not represent an issue. At the same time, expert terminologists also identify those entities that instead may represent an issue, because not present in both. Furthermore, by placing equivalence relationships between the graphs, some further issue related to knowledge organization – different hierarchies, inverted parenthood, etc. – can be recognized and solved by editing either ICD11 or SNOMED-CT.

However, this is only a first experiment to set guidelines and understand the amount of work needed for the whole Common Ontology project. In fact, mapping guidelines are being developed and will be made available in due course to provide a common framework to the experts possibly involved in further steps.

The total number of entities involved in the three ICD11 chapters currently under examination is 2796 on a total of 32995 entities in the whole ICD11 (at the time of writing). The percentage of entities being analysed in this preliminary work is thus about 8.5%, a proportion we considered adequate to for a feasibility study.

The final output of this preliminary step will help in setting an agenda for the next few years, whose aim will be to develop an effective tool for maintaining semantic interoperability between SNOMED-CT-annotated clinical records and ICD11-classified data initially, then involving also the other WHO classifications, namely the International Classification of Functioning, Disability and Health (ICF) [10] and the forthcoming International Classification of Health Interventions (ICHI) [11].

ACKNOWLEDGMENT

This work was supported by the World Health Organization (WHO) and the International Health Terminology Standards Development Organization (IHTSDO) through their Joint Advisory Group (JAG).

O.V. work is funded by the Italian WHO-FIC Collaborating Center at the Central Health Directorate, Classification Area, Friuli Venezia Giulia Region (Udine).

We wish to thank our JAG partners: Hazel Brear, Sukil Kim and Harold Solbrig.

REFERENCES