Title: Publishing a Multilingual Medical Terminology According to Terminology Standards and Linked Data Principles.

Authors: Maxime WARNIER[1], Joseph ROUMIER[2], Marc JAMOULLE[3], Elena CARDILLO[4], Robert VANDER STICHELE[5], LAURENT ROMARY[6]
[1]UCL-CENTAL, Belgium, student in Master degree, MaximeWarnier@gmail.com
[2]CETIC & Heymans Institute, Belgium, Joseph.Roumier@cetic.be
[3]IRSS-UCL, Belgium, Marc.Jamoulle@uclouvain.be
[4]eHealth-FBK, Italy, Cardillo@fbk.eu
[5]Inria & HUB, France, Laurent.Romary@inria.fr
[6]Heymans Institute, University of Ghent, Belgium, Robert.VanderStichele@ugent.be

Abstract: The article gives an overview of the results obtained by converting a focused medical resource into RDF triples and linking it with other reference resources.

Keywords: primary care, multilingual terminology, ontology, Linked Data.

The data source is a *multilingual medical database* manually created by Mr Marc Jamoulle, a Belgian general practitioner with a long experience in classification and terminologies for general practitioners. It consists of one hundred and seventy-three French terms identified in a guideline concerning heart failure [1], intended for family physicians and published by the Société Scientifique de Médecine Générale (SSMG, Belgium). This resource is a first step towards the creation of a Medical Reference Terminology [2]. Due to the readership of the publication, those terms are often distinct from the ones used in the common language, as well as from those – even more technical – used by the specialized cardiologists; this situation clearly proofs that, because of the diversity of the terms, interoperability is sometimes hard to preserve. All of them have also been translated into English and the corresponding concepts have been retrieved in four widely recognized and used international classifications (UMLS[1], SNOMED-CT[2], ICD10[3] and ICPC2[4]). This allowed to collect their lexical representations and the corresponding internal codes (if available), along with definitions and other useful information. This preliminary work was part of the Meriterm project.

The major part of the information present in the database was converted into nearly **13.000 RDF triples** (serialized in an RDF/XML document) using the *Jena*[5] API for Java. This task was performed in a fully automated manner. The resulting resource can be found at the following URI: http://meriterm.org/heartfailure/heartfailure.rdf[6].

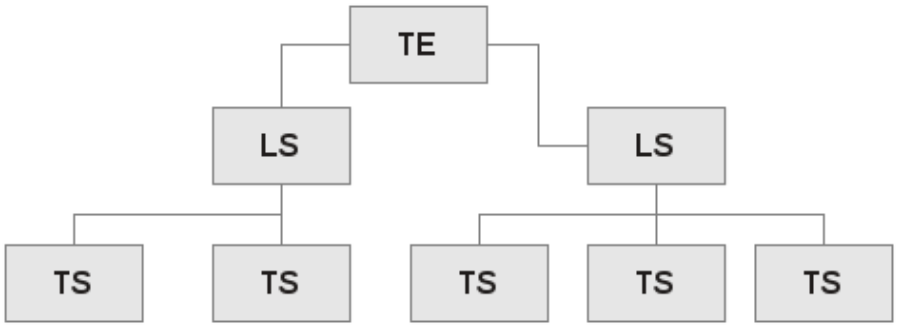
The resulting resource is a focused terminology that contains well-defined concepts, linked to reference resources of the field, with the associated term(s) in French and English (Dutch and Italian are next to be added). It was therefore mandatory to efficiently handle its multilingual nature. Moreover, as already mentioned, it was necessary to distinguish between the various words or phrases coexisting in a same language[7] and to specify the status of each of them (standardized, admitted or preferred).

This leads to a model based on the Terminological Markup Framework (TMF) [3], standardized by the ISO committee (ISO16642 [4]). This model will be described more in

1 http://www.nlm.nih.gov/research/umls/
2 http://www.ihtsdo.org/snomed-ct/
3 http://www.who.int/classifications/icd/en/
4 http://www.who.int/classifications/icd/adaptations/icpc2/en/index.html
5 http://jena.apache.org/
6 Using cURL:
```
curl -L -H "Accept: application/rdf+xml"
http://meriterm.org/heartfailure/heartfailure.rdf
```
7 For example, the diseases commonly referred to as *cancers* in common English are known as *malignant neoplasms* by the specialists.

details in a further publication. The main components are called *Terminological Entry* (TE), *Language Section* (LS) and *Term Section* (TS). Thus, concepts (at a higher level) and terms are kept separated, but are strongly linked, since each TE can contain an unlimited amount of LS (currently two, for French and English) and, in the same way, each LS can have an undefined number of TS (one for each lexical representation of the concept). Obviously, in order to provide useful information, they must be described using several properties. Those properties link the components to data categories (definition, domain expert...), which, in turn, may be linked to other categories (e.g. a definition can be associated to its source). Similarly to Lemon, the data categories chosen come from ISOcat (ISO12620 [5]). The whole OWL vocabulary is declared at http://meriterm.org/heartfailure/vocabulary.owl.

For convenience, direct access to every component is also provided by alternative, more readable (and elegant) URIs[8]. For instance, accessing http://meriterm.org/heartfailure/25/en/preferred/ will return an RDF/XML document containing only the few triples that are related to http://meriterm.org/heartfailure/heartfailure.rdf#25_TS_EN_PREF (that is, the *preferred term* of the English language section of the entry whose identifier is 25).[9] On top of it, this spares some bandwidth and processing resources, which is relevant, especially for mobile devices.



A very simplified view of the model

Likewise most available semantic linguistic resources, concepts (i.e. terminological entries) are represented as classes (with the standardized lexical representation as a label annotation), while the lexical representations (i.e. instances of Term Section) are represented as individuals. Whenever possible, the former are declared *equivalent classes* to classes found in ontologies present on the NCBO BioPortal[10] and in the Data Hub, which have a high visibility (ICD10 is supervised by the World Health Organization and SNOMED-CT is maintained by the International Health Terminology Standards Development Organisation). Although this choice may not be exclusive, *equivalentClass* property were preferred over *sameAs* property, even if the *sameAs* property is more common, because an identity link may induce undesired affirmations and the task of the reasoners could become much harder [6]. However, a given concept does not always perfectly match the ones (that can be defined as *unionOf* concepts) found in the targeted ontologies or terminologies. This can be considered a problem as the formalism used might be too strong.

Finally, the dataset can be queried using SPARQL; for this purpose, a SPARQL endpoint is provided at **http://meriterm.org:8081/openrdf-**

8 Solution inspired by the explanations given by OpenLink Software (http://virtuoso.openlinksw.com/whitepapers/VirtDeployingLinkedDataGuide_Introduction.html#mozTocId502192).

9 Or http://meriterm.org/heartfailure/25/fr/ for its French language section, or simply http://meriterm.org/heartfailure/25/ for the terminological entry itself.

10 http://bioportal.bioontology.org/

sesame/repositories/meriterm. Here are some examples of queries:

1. Find all the different lexical representations for a concept (e.g. http://meriterm.org/heartfailure/heartfailure.rdf#1) :

```
PREFIX voc: <http://meriterm.org/heartfailure/vocabulary.owl#>
PREFIX hf: <http://meriterm.org/heartfailure/heartfailure.rdf#>

SELECT DISTINCT ?term

WHERE {
       hf:1 voc:hasLS ?ls .
       ?ls voc:hasTS ?ts .
       ?ts voc:isTerm ?term
}
```

2. Find all the concepts that are present in UMLS, but not in ICPC2:

```
PREFIX voc: <http://meriterm.org/heartfailure/vocabulary.owl#>
PREFIX hf: <http://meriterm.org/heartfailure/heartfailure.rdf#>

SELECT ?concept ?code_umls

WHERE {
       ?concept voc:hasClassificationCode ?classification_code_umls .
       ?classification_code_umls voc:hasOriginatingDatabaseName hf:UMLS .
       ?classification_code_umls voc:hasValue ?code_umls

       OPTIONAL {
              ?concept voc:hasClassificationCode ?classification_code_icpc .
              ?classification_code_icpc voc:hasOriginatingDatabaseName hf:ICPC2
       }

       FILTER(! bound(?classification_code_icpc))
}
```

Aknowledgement:

Part of the work was done within the Meriterm project, that started in 2011 between the Centre d'Expertise en Technologies de l'Information et de la Communication, the Heymans Institute of Pharmacology, the University of Ghent, the Université de Louvain and the Fondazione Bruno Kessler, and aims at creating a framework for publishing and editing medical multilingual linguistic resources using semantic web concepts and tools as well as existing metadata, terminology and lexicon standards.

References:

[1] P. Van Royen, P. Chevalier, G. Dekeulenaer, M. Goossens, P. Koeck, M. Vanhalewyn, P. Van den Heuvel. Recommandation de Bonne Pratique. Insuffisance cardiaque, 2011. http://www.ssmg.be/images/ssmg/files/Recommandations_de_bonne_pratique/RBP_insuffisance_cardiaque.pdf

[2] J. Roumier, R. Vander Stichele, L. Romary, and E. Cardillo, “Approach to the Creation of a Multilingual, Medical Interface Terminology,” Nov. 2011. http://hal.inria.fr/hal-00646223_v1/

[3] L. Romary, "An abstract model for the representation of multilingual terminological data: TMF - Terminological Markup Framework", Proc. TAMA 2001 – http://hal.inria.fr/inria-00100405

[4] ISO 16642:2003, Computer applications in terminology – Terminological markup framework (TMF), 2003.

[5] ISO 12620:1999, Computer applications in terminology – Data categories, 1999.

[6] H. Halpin, I. Herman, P. J. Hayes. When owl:sameAs isn't the Same: An Analysis of Identity Links on the Semantic Web, 2010.