Approach to the Creation of a Multilingual, Medical Interface Terminology

Joseph Roumier
Heymans Institute of Pharmacology / Ghent, Belgium
CETIC / Charleroi, Belgium
LT3 / Ghent, Belgium
Joseph.Roumier@cetic.be

Robert Vander Stichele
Heymans Institute of Pharmacology / Ghent, Belgium
Robert.VanderStichele@ugent.be

Laurent Romary
INRIA & HUB-IDSL / Paris, France
Laurent.Romary@inria.fr

Elena Cardillo
Fondazione Bruno Kessler / Trento, Italy
cardillo@fbk.eu

Abstract

Health care professionals experience difficulties in the correct medical registration of clinical work and in the efficient searching for answers to clinical questions. These difficulties arise often from a deficient interface between human and machine language. Terminological solutions are often naive attempts to standardize language and terms, with conceptual systems, which may overwhelm the users by their complexity, or be too restrictive to represent crucial details. Moreover, local, professional and cultural differences in vernacular expression are often not represented.
We must take into account vocabulary differences between Specialists and General Practitioners talking about the same medical fact. There are even more differences between the languages of patients and physicians. Also, the vocabulary being used evolves over time and space and many local expressions exist to designate the same diseases or body parts.

In order to cope with this heterogeneity in a(n) (semi)automated manner, in fact to perform the task of the doctor being the interface between the medical wor(l)d and the lay person wor(l)ds, Natural Language Processing is necessary. Even though some mechanisms exist, the effort to maintain a central and evolving multilingual terminology containing all the linguistic complexities and the local lexical variants for a concept is daunting.

That is why we propose a terminological system that contains two types of domain-specific resources.

- First, a reference terminology [Rosenbloom et al., 2006], [Rosenbloom et al., 2008], multidisciplinary, multilingual but containing only the reference concepts and their standardized as well as local lexical representations. These reference concepts are linked to nomenclatures, such as SNOMED-CT[1], or bibliographic thesauri such as Medical Subject Headings (MeSH), or to international classifications .
- The second resource, is a series of specific, lexical and often monolingual "end-user terminologies" that must be linked to the multilingual reference terminology. These lexicons can be linked to Natural Language Processing applications, and be oriented to patients or to professionals (e.g. local nomenclatures, coding systems, etc.).

We propose a dual mechanism to link the first type of resource – the reference terminology - and the second type of resources – the end-user monolingual terminologies:

[1] http://www.ihtsdo.org/snomed-ct/

- The concept in the reference terminology is linked to the sense part of a lexical resource. This mechanism preserves the conceptual integrity and is language independent.
- The alignment of the lexical representation in a specific language of the concept with the corresponding lemma in a lexicon of that language. This is language dependent.

The reference terminology is created using the association of the Terminological Markup Framework (TMF) [ISO 16642, 2003], [Romary, 2010]) as the meta-model and a carefully chosen subset of the data-categories found on the ISOcat.org[2] platform [ISO 12620, 1999]. The resulting Terminological Markup Language (TML) (see Figure 1) is serialized using RelaxNG[3]. For this part of the work, we were inspired by the TML created by the TermSciences[4] project [Khayari et al., 2006].

To ensure semantic interoperability, we linked this new resource first of all with SNOMED CT[5], used as the backbone of the reference terminology. In addition, we link the concepts to MeSH [Lipcomb, 2000] and to a series of other external classifications, such as the International Classification of Dis-

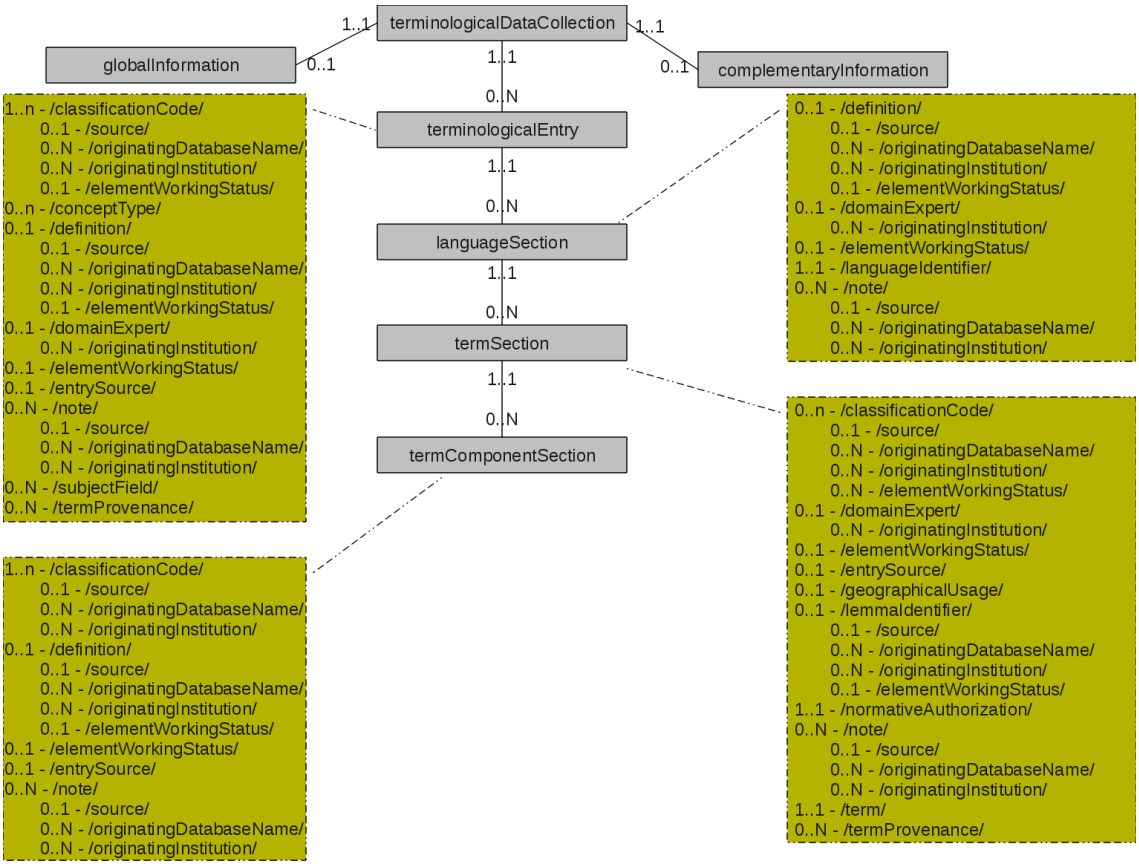

Figure 1: TMF and Data Categories for the main levels

eases (ICD) [Slee et al., 1978] [ICD-10, 2010]; the International Classification for Primary Care – (ICPC) [ICPC-2-R, 2005]; and LOCAS [Jamoulle & Roland, 1993] a Primary Care oriented resource.

To represent the end-user terminologies, our proposal is to use the Lexical Markup Framework(LMF) [ISO 24613, 2008], [Romary, 2010] presented in Figure 2, since it is conceived to deal with linguistic complexities, and uses the same source - ISOCat.org - for linguistic Data Categories. LMF provides

[2] http://www.isocat.org/
[3] http://www.relaxng.org/
[4] http://www.termsciences.fr/
[5] http://www.ihtsdo.org/snomed-ct/

guidelines on how to link its entries with TMF and other concept based representation systems. Finally the lexical meta-model contains a mechanism to deal with multiple senses, and possibilities to link with lexi-ontological resources for patients [Cardillo, 2011].

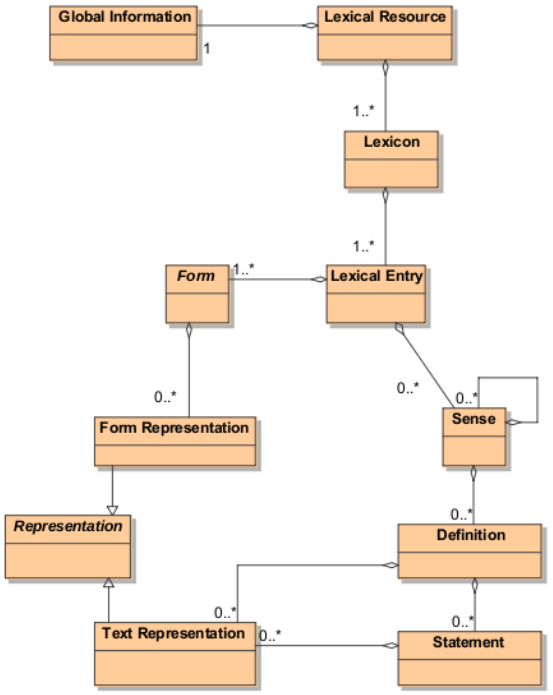

Figure 2: LMF Meta model – core package – (from ISO 24613:2008)

In addition to this, we plan to publish these resources online following the Linked Data [Bizer et al., 2010] principles produced by the Semantic Web Initiative of the World Wide Web Consortium (W3C)[6]. Guidance towards this goal is provided for TMF, LMF and ISOCat.org, respectively in [Romary et al., 2006], [Hayashi et al., 2011] and [Kemps-Snijders et al., 2009].

In conclusion the work relies on three pillars: first the existence of an ISO standard to develop a model for a multilingual reference terminology that take into account the diversity of terminology sources while preserving interoperability and sustainability. Second, the existence of an ISO standard to develop models for mono- (and multi-) lingual lexical end-user terminologies, that tap into the existing body of language-specific linguistic resources. Third, the existence of W3C standards for the publication of resources as Linked Open Data. The work is ongoing and currently being reviewed by the health and environment department of the Belgian government.

References

[Bizer et al., 2010] Bizer, C. and Heath, T. and Berners-Lee, T., Linked data-the story so far, sbc, S.9, 2010

[Cardillo, 2011] Cardillo E. A lexi-ontological resource for consumer healthcare. The Italian Consumer Medical Vocabulary. [Doctoral Thesis]. Fondatione Bruno Kessler. April 2011.

[6] http://w3.org/

[Hayashi et al., 2011] Hayashi, Y. and Declerck, T. and Calzolari, N. and Monachini, M. and Soria, C. and Buitelaar, P., Language Service Ontology, The Language Grid, S.85-100, 2011

[ICPC-2-R, 2005] ICPC-2-R - Wonca (World Organisation of Family Doctors) ICPC-2-R, International Classification of Primary Care (revised 2nd Ed). OUP. 2005

[Slee et al., 1978] Slee, V. N. and others,, The International Classification of Diseases: ninth revision (ICD-9), Annals of internal medicine, S.424, 1978

[ICD-10, 2010] ICD-10 -International Statistical Classification of Diseases and Related Health Problems. 10th Revision. Version for 2007.Tabular List of inclusions and four-character subcategories-http://apps.who.int/classifications/apps/icd/icd10online/

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

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

[ISO 24613, 2008] ISO 24613:2008, Language resource management — Lexical markup framework (LMF)

[Jamoulle & Roland, 1993] « LOCAS-CISP, logiciel de codage et d'acquisition de synonymes pour la Classification Internationale des Soins Primaires » (Jamoulle M, Roland M Fédération des Maisons Médicales, Bruxelles), 1993.

[Kemps-Snijders et al., 2009] Kemps-Snijders, M. and Windhouwer, M. and Wittenburg, P. and Wright, S. E., ISOcat: remodelling metadata for language resources, International Journal of Metadata, Semantics and Ontologies, S.261-276, 2009

[Khayari et al., 2006] Khayari, Majid and Schneider, Stéphane and Kramer, Isabelle and Romary, Laurent, Unification of multi-lingual scientific terminological resources using the ISO 16642 standard. The TermSciences initiative., 2006, http://hal.archives-ouvertes.fr/hal-00022424

[Lipcomb, 2000] Lipscomb, C. E., Medical subject headings (MeSH), Bulletin of the Medical Library Association, S.265, 2000

[Romary et al., 2006] Romary, Laurent and Kramer, Isabelle and Alt, Susanne and Roumier, Joseph, Gestion de données terminologiques : principes, modèles, méthodes, S.13, 2006, http://hal.archives-ouvertes.fr/hal-00096910

[Romary, 2010] Romary, Laurent, Standardization of the formal representation of lexical information for NLP, 2010, http://hal.inria.fr/hal-00436328

[Rosenbloom et al., 2006] Rosenbloom ST; Miller RA, Johnson KB, Elkin PI, Brown SH. Interface terminologies: Facilitating direct entry of clinical data into electronic health record systems. J Am Med Inform Assoc 2006;13:277-288.

[Rosenbloom et al., 2008] Rosenbloom ST, Miller RA, Johnson KB, Elkin PI, Brown SH. A model for evaluating interface terminologies. J am Med Inform Assoc 2008;15:65-76.