Approach to building multilingual terminological resources using the ISO-standards Terminological Markup Framework (TMF) and Lexical Markup Framework (LMF).

A use case for heart failure.

Joseph Roumier
Independent Consultant

Annex 2
to SHN Work Package 3
Deliverable D3.3

Final Version, May 29, 2015
Deliverable: Annex 2 to SHN WP3 Deliverable D3.3

Publishable summary: In this Annex, the author describes the existing ISO-standards for multilingual terminologies, both lexical (Lexical Markup Framework LMF) and conceptual (Terminological Markup TMF; ISO-CAT (the international resource of data categories and domain values for linguistic resources); and Linked Open Data as a technology to publish semantic interoperability resources on the semantic web. The structure is described of a proof-of-concept multilingual interface terminology for heart failure.

Status: Final draft
Version: 1.3
Public: Yes
Deadline: May 29, 2015

Contact: Joseph.Roumier, joseph.roumier@fdn.fr
Robert Vander Stichele robert.vanderstichele@ugent.be

Editors: Joseph Roumier

Table of Content

Introduction .................................................................................................................................................. 3
Description of the four pillars of the approach........................................................................................................... 4
1. Terminological Markup Framework (TMF) for multilingual reference terminologies ........................................ 4
2. Lexical Markup Framework (LMF) for mono- and multilingual lexical terminologies ........................................ 5
3. International Catalogue of data categories and domain values for linguistic resources .................................... 5
4. Semantic web technology to publish terminological resources on the World Wide Web: Linked open Data ........................................................................................................................................... 6
A Use Case: Interface terminology for Heart Failure ................................................................................................. 8
Conclusion ....................................................................................................................................................... 9

Note:
This annex was commissioned by Prof. Dr. R. Vander Stichele, Workpackage Leader of SemanticHealth Net WP3, to Joseph Roumier, independent Consultant, joseph.roumier@fdn.fr
Introduction

The complexity of medical linguistic resources requires more than in other fields a robust and standardized architecture and methodology to create and maintain them. With the advent of internet, we are able to communicate crucial information to every connected part of the planet. For it to be useful requires not only multilingualism but also semantic interoperability and interlinking of a vast variety of linguistic resources.

This annex describes an approach that relies on four pillars:

- first the existence of an ISO standard to develop models for multilingual reference terminologies 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 lexicons, that tap into the existing body of language-specific linguistic resources.
- Third, the use of a central catalogue of data categories and domain values for terminological resources (ISO-Cat\(^1\))
- Fourth, the existence of a W3C\(^2\) standards for the publication of semantic data on the Internet.

This approach has been developed in the past years and applied to proof-of-concept projects, namely the creation of Reference Terminology for collections of concepts extracted from a Heart Failure guideline\(^3\), and a multilingual glossary of technical and popular terms in patient package inserts for drug information\(^4\).

---

1 ISO TC37 Data Category Registry (DCR) - [http://www.isocat.org/](http://www.isocat.org/)
3 Cardillo, Elena and Roumier, Joseph and Jamoulle, Marc and Vander Stichele, Robert and others, Using ISO and Semantic Web standards for creating a Multilingual Medical Interface Terminology: A use case for Heart Failure., 2013
4 Stichele, R. Vander, Multilingual glossary of technical and popular medical terms in nine European languages, 1995
Description of the four pillars of the approach

1. Terminological Markup Framework (TMF) for multilingual reference terminologies

TMF stands for Terminological Markup Framework and provides a Meta-Model for onomasiological, multilingual, multilevel terminologies. It can be used to manage collections of concepts, relevant to a specific domain in medicine (e.g. heart failure, drug information). It can also be used to manage national sub-sets of international terminologies. Finally, it can be used to master the multilingual management of international classifications (e.g. International Classification of Primary Care (ICPC), Anatomical Therapeutic Classification (ATC)). It is based on UML and XML, but any database backend can be used for the terminologies, including triplestores (semantic web oriented databases).

Figure 1: Terminological Markup Framework (TMF) meta-model for multilingual, multilevel terminologies (Romary, L.)

---

5 ISO 16642, Computer applications in terminology - Terminological markup framework (TMF), 2003
6 Romary, Laurent,  Standardization of the formal representation of lexical information for NLP, 2010, [http://hal.inria.fr/hal-00436328](http://hal.inria.fr/hal-00436328)
2. Lexical Markup Framework (LMF) for mono- and multilingual lexical terminologies

LMF stands for Lexical Markup Framework and provides a Meta-Model for mono and multilingual semasiological lexicons. This ISO standard is also based on UML and XML. The core components of the LMF meta-model can be extended to support multilingual lexicons by the NLP (Natural Language Processing) multilingual notations extension, using the “sense” component as a link between several languages. The lexicons are necessary to cover the large number of lexical representations of a concept as well as the variants.

![Diagram of the LMF meta-model](image)

Figure 2: Lexical Markup Framework (LMF) meta-model (ISO 24613:2008)

3. International Catalogue of data categories and domain values for linguistic resources

Following the guidelines of the ISO 12620:2009, the TC37 Data Category Registry (DCR) provides in the form of a database the required data categories for language resources. It is available on the ISOCat plateforme. Linguistic Data Categories (/part of speech/, /gender/, /language/, etc.) are chosen according the needs of a
given model and hooked on the meta-models of the terminologies and lexicons.

The availability of the data categories in the Web Ontology Language [OWL] allows for easier integration in semantic-web oriented information systems.

### 4. Semantic web technology to publish terminological resources on the World Wide Web: Linked open Data

Following the principles of the semantic web and the goal of using the web not only for websites and webpages but also for data, the Linked Data initiative has been successfully used to publish and interlink an increasing number of medical and non-medical resources. OBOFoundry and BioPortal which publish currently hundreds of ontologies, terminologies, nomenclatures, along with the mappings/links between them are examples of this success. The pink disks in Figure show the life science resources published as linked open data, as well as the links with other sources, from the life science domain or other domains.

Using the concepts and tools of the semantic web to provide terminologies, lexicons, nomenclatures and other medical resources as linked open data augments the chances for them to be used, reviewed and enhanced.

---

11 Horrocks, Ian and Patel-Schneider, Peter F. and Harmelen, Frank van, From $\mathcal{SHIQ}$ and RDF to OWL: The Making of a Web Ontology Language, J. of Web Semantics, S.7-26, 2003
12 Horrocks, I., Semantic web: the story so far, Proceedings of the 2007 international cross-disciplinary workshop on Web accessibility (W4A), S.120-125, 2007
13 Bizer, C. and Heath, T. and Berners-Lee, T., Linked data-the story so far, sgc, S.9, 2010
Figure 3: Linking Open Data cloud diagram 2014, by Max Schmachtenberg, Christian Bizer, Anja Jentzsch and Richard Cyganiak. http://lod-cloud.net/
A Use Case: Interface terminology for Heart Failure

In this use case, a model for a multilingual interface terminology was created to manage medical concepts in a bilingual Belgian guideline for GPs\textsuperscript{16}. The interface terminology consists of monolingual lexicons in LMF (with Lemon\textsuperscript{17} elements) and a multilingual reference terminology in TMF. The terms in the French version of the heart failure guideline were extracted manually and entered into the terminological databases with a concept definition. Then the selected concepts were coded first in ICPC-2, then ICD-10\textsuperscript{18} by a physician, an expert in classification systems. In addition, mappings were sought with SNOMED-CT\textsuperscript{19} and UMLS\textsuperscript{20} concepts, using the UMLS SNOMED-CT browser\textsuperscript{21}. Finally, Dutch, English, and Italian translations of terms were added, as well as lay term terms for the concepts in the different languages.

The methodology to create this resource was to:

- define the resource and its required information facets
- decide whether it was multilingual, and multi-level (e.g. Medical jargon and lay man language)
- Decide whether connections are needed to other resources (ontologies, other terminologies, lexicons)
- Identify corresponding data categories in a linguistic data category repository, (ISOcat) and create specific data categories to the new resource, if needed, after consultation with ISOcat
- analyse the general TMF meta-model and identify the correct node to attach the data categories to into a resource-specific data model (see figure 3 for an example)
- serialize the model as a XML repository
- populate the databases with terms (e.g. extracted from a guideline)
- export to a triplet database
- Publish as Linked open Data.

\textsuperscript{17} The Lexicon Model for Ontologies - http://www.lemon-model.net/
\textsuperscript{18} International Classification of Diseases – http://www.who.int/classifications/icd/en/
\textsuperscript{19} SNOMED-CT - multilingual clinical healthcare terminology – http://www.ihtsdo.org/
Conclusion

This annex describes how stable standards and methodologies are applied to create perennial and interconnected medical linguistic resources based on ISO standards and W3C recommendations. It is useful to harness the size and assure the maintainability of reference terminologies, by assuring external mapping and linking to autonomously evolving nomenclatures, classifications, thesauri, lexicons and glossaries. terminologies, in a multilingual and multidisciplinary approach. Publication of this resource as Linked Data will assure connectivity of the reference terminology with the rapidly evolving collection of resources in the semantic web.