



An Online Modeling Tool  
for building  
Crowd-sourced and Multilingual  
Terminological Resources  
in Semantic Health

Olivier Latignies

Consultant R&D Management of web based software solutions

**Annex 5**  
**to SHN Work Package 3**  
**Deliverable D3.3 of March 31, 2015 .**

## Final version, March 29, 2015

### Document description

Deliverable:	Annex 5 to SHN WP3 Deliverable D3.3
Publishable summary:	In this Annex, the author describes the architecture of an interactive website for the collaborative, crowd-sourced production and maintenance of interface terminologies as resources for semantic interoperability. The architecture has a dual structure, with a server for the data in an object-oriented and in a semantic framework, and permits the publication of the resource in Linked Open Data
Status:	Final draft
Version:	1.5
Public:	<input checked="" type="checkbox"/> Yes
Deadline:	March 29, 2015
Contact:	Olivier Latignies <a href="mailto:ol@two4tek.eu">ol@two4tek.eu</a> Robert Vander Stichele <a href="mailto:robert.vanderstichele@ugent.be">robert.vanderstichele@ugent.be</a>
Editors:	Olivier Latignies

### Table of Content

Introduction.....	4
Basic requirements .....	5
Use of standards for multilingual terminologies .....	5
Ability to use formats and tools to build and use semantic web resources .....	5
Ability to combine formats and views of terminological resources.....	5
Ability to creating an information rich environment .....	6
Multilingual requirements .....	6
Integration of international reference terminologies.....	7
Workflow requirements .....	8
Collaboration .....	8
Revision cycle .....	8
Model management .....	8
Concept management.....	8
Translation management .....	8
Release management.....	9
Access control.....	9
Performance.....	9
Feedback loop.....	9
The dual architecture of the application .....	10

Object-oriented database server.....	10
Ontology server.....	10
The dual model .....	11
A web-based application .....	11
Final remarks .....	12

Note :

This annex was commissioned by Prof. Dr. R. Vander Stichele, Workpackage Leader of SemanticHealth Net WP3, to Olivier Latignies, Crisnet ASBL, Coordination, Recherche et traitement de l'Information en Soins de santé primaires, NETWORK, Rue des Aulnois, 25 - B-7090 Hennuyères (Belgique)

Email: [ol@two4tek.eu](mailto:ol@two4tek.eu)

# Introduction

This paper describes an on-line modeling tool to create and maintain publically available semantic terminological resources for the creation and maintenance of translations of reference terminologies, and to publish them on the web in Linked-Open Data.

The use-cases for this approach will be the multilingual management of a small domain-specific conceptual reference terminology for primary care ((n=5000 to 10.000 concepts), encapsulating the International Classification of Primary Care (ICPC), linked to domain-specific unilingual lexical dictionaries and to the bigger international reference terminologies, such as SNOMED-CT, ICD, LOINC, UMLS, MeSH,

This application has 3 basic functionalities:

- Selection of candidate concepts and preferred terms
- Translation of selected preferred terms to other languages
- Mapping of selected concepts to international reference terminologies

# Basic requirements

## Use of standards for multilingual terminologies

LMF<sup>1</sup> (Lexical Markup Language) and TMF<sup>2</sup> (Terminology Markup Language) are two standards for structuring Multilingual Terminologies. LMF is suited for lexical terminologies (e.g. dictionaries), while TMF is used for reference terminologies for machine language and coding systems.

These 2 standards are quite complex to implement. They are mainly used to provide a mechanism to store resources and to distribute them but not to be used as a database. The manipulation of the resource will be deferred to a specialized tool providing a human interface usable by the researchers.

## Ability to use formats and tools to build and use semantic web resources

RDF<sup>3</sup> (Resource Description Format) and OWL<sup>4</sup> (Web Ontology Format) are formats to describe Web Semantic Resources also known as Ontologies, but can also be used for describing classifications or reference terminologies, and to express the LMF/TMF formats.

A well-known and widely used tool to build ontologies and reference terminologies is Protégé<sup>5</sup>, now used in the construction of the next version of the International Classification of Diseases (ICD-11).

Many other ontological tools are available for free or commercially, each other with more or less refinements and add-ons. A list is available on Wikipedia [http://en.wikipedia.org/wiki/Ontology\\_editor](http://en.wikipedia.org/wiki/Ontology_editor)).

## Ability to combine formats and views of terminological resources.

The current ontology editors are aimed at experts in semantics and ontologies, understanding the complexity of semantic graphs and the related standards.

In our case, the approach must be task-oriented towards the building of a terminological resource. We must build a way to integrate a guided workflow, with a step by step procedure to encode new terms and their relations to other terms, to classifications and other terminologies.

---

<sup>1</sup> "Lexical Markup Framework (LMF)." 2006. 26 Feb. 2015 <<http://www.lexicalmarkupframework.org/>>

<sup>2</sup> "TMF Webpage - Loria." 26 Feb. 2015 <<http://www.loria.fr/projets/TMF/>>

<sup>3</sup> "RDF - Semantic Web Standards." 26 Feb. 2015 <<http://www.w3.org/RDF/>>

<sup>4</sup> "OWL - Semantic Web Standards." 2004. 26 Feb. 2015 <<http://www.w3.org/2004/OWL/>>

<sup>5</sup> "protégé." 2002. 26 Feb. 2015 <<http://protege.stanford.edu/>>

To be able to be crowd-sourced and to be used by physicians, the human interface must be easy to use and should not require a long learning curve. The contributors will be physicians with a limited amount of time for training to use a tool that they will use only a couple of hours per month.

To make this come true, an environment must be created, where these resources and mutual mappings are available and accessible in different formats for different views and user interfaces, such as CVS (Comma Separated Value) for spreadsheet view or database view, or RDF, for views as ontological graphs.

## **Ability to creating an information rich environment**

In the Health sector, a number of reference terminologies and classifications exist that can be used to guide the user in the selection of concepts, or in creating links from words and phrases to concepts, and a host of international reference terminologies.

The idea is to bring these existing terminologies and their mutual mappings in a single environment, to support a community of primary care physicians and allied health personal, working on a domain-specific core set reference terminology.

If we use the transcoding (or mapping) tables of each of these resources, we can ease the search in the larger terminologies, to facilitate the identification of the correct mappings.

We can building on previous experience users already have in working with smaller classifications, such as the International Classification of Primary Care (ICPC) or International Classification for Nursing Practice (ICPN).

These simple and ergonomic axial classifications can help to identify small sets of possibly relevant codes in SNOMED, ICD, ICF, UMLS, MESH or ICD codes, by using the mappings that already exist between these two resources.

## **Multilingual requirements**

As the purpose of this tool is to build multilingual resources, the complexity of the languages must be taken into account when building the tool interface, by providing the multilingual character sets and multilingual management of the interface.

## Integration of international reference terminologies.

Browsers to international reference terminologies may be very different. Simple tables (ICPC<sup>6</sup>), hierarchical graphs (ICD<sup>7</sup>), complex relational graphs in RDF/OWL may confront the user with a bewildering diversity.

The tool will focus the user on his task by presenting all the resources in a systematic way.

If the user want to link a new concept in the domain-specific core set reference terminology, to an international reference terminology, it must be done naturally by selecting first the target reference terminology.

Then the user will be able to use the ICPC, and its mappings to the targeted resource, to be presented with a selection of codes, linked with the ICPC code.

---

<sup>6</sup> "WICC - Global Family Doctor - Wonca Online." 2012. 27 Feb. 2015  
<<http://www.globalfamilydoctor.com/groups/WorkingParties/wicc.aspx>>

<sup>7</sup> "WHO | International Classification of Diseases (ICD)." 27 Feb. 2015 <<http://www.who.int/whosis/icd10/>>

# Workflow requirements

## Collaboration

As in all crowd-sourced project, collaboration is the key to success.

Each step of the process must integrate a way to collaborate between experts by opening a conversation linked to a resource (or a part of it) and provide the tools (status, comments,...) to settle the case.

All the activities must send notifications to all the parties to require their attention on a topic they are involved in. This will create a dynamic environment increasing the involvement of the users. Editors will receive rights according to the functionalities in which they are involved (concept selection, translation of preferred terms, mapping of concepts).

## Revision cycle

Every steps of the revision cycle will be constrained by peer review. The tool will control all the steps to ensure that every information is accurate and validated before using it.

History, comments and status will be part of the cycle to trace any problems and to be able to identify the source, to resolve it with the help of the user how as created or modified it and so to speed up the process.

## Model management

A new model must be built to connect the standards that will be used.

This model will change over time to accommodate the findings and optimisations the researchers will discover building the resource.

## Concept management

New concepts will be created or modified throughout the use of the tool. These changes must be validated by experts before being able to use the new concepts.

## Translation management

Translations will be done by users with deep knowledge of the terms describing a concept and it's way the words are put together.

## **Release management**

To be able to release stable versions, a revision cycle must be put in place.

Freezing all new additions and beginning a full revision of the current terms is mandatory before releasing a stable version. These steps will be integrated in the tool itself to streamlined the release process.

## **Access control**

All actions will be constrained by an access model allowing only validated user to take action on the information they are allowed to access.

All information should be controlled by this access model for reading or updating it's value.

## **Performance**

Performance is critical in the feeling the user will have regarding the tool and the effort he will put into it. If the tool is responsive and does not let the user wait to much between each actions, it will be easier for the user to use the tool and so produce a better job and a better information quality.

## **Feedback loop**

This tool is meant to be used in a feedback loop with other software using the resource to validate it in real world examples.

Each finding should be used to improve the resource through the revision cycle.

# The dual architecture of the application

## Object-oriented database server

On the one hand, we manipulate information models in an object database abstraction database model, stored in an SQL-database and accessible to end-users.

## Ontology server

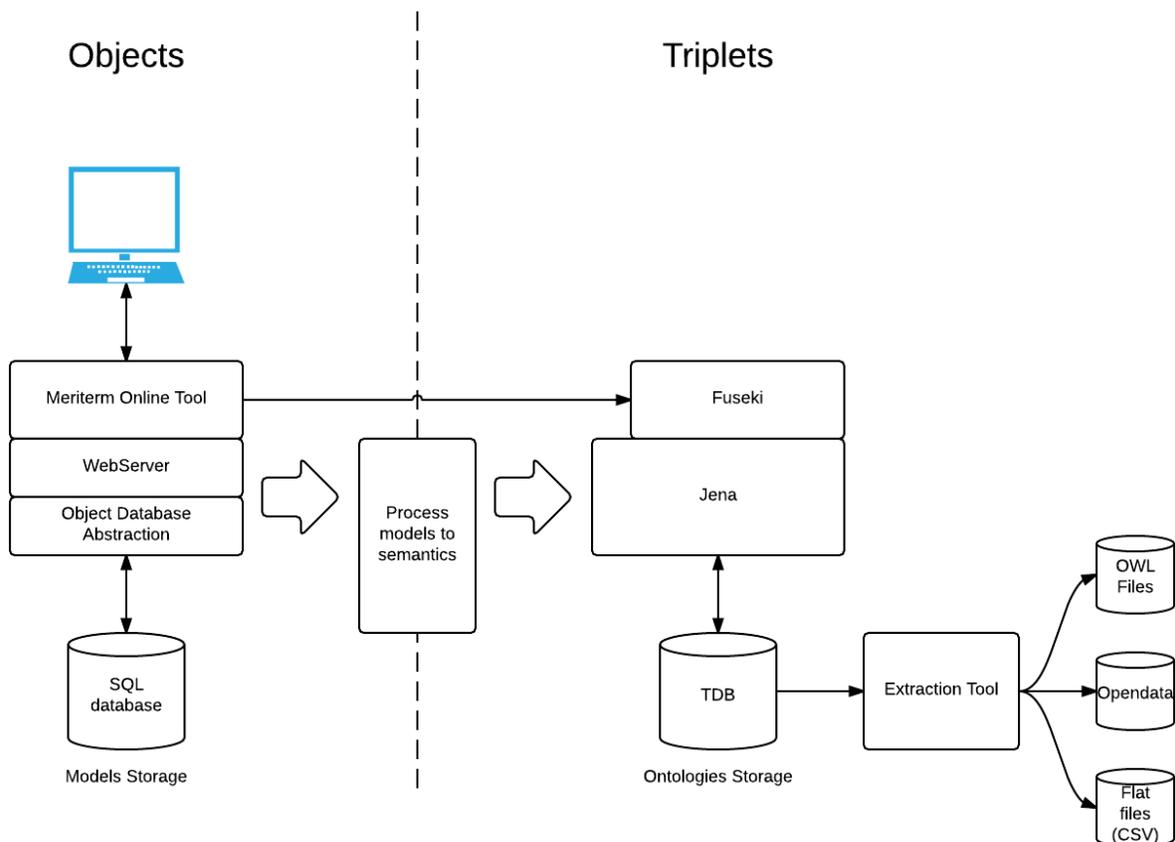
On the other hand, we manipulate triplets and semantic models. Here, the best language for this purpose is SPARQL. Sparql is a Query Language used to manipulate triplets databases (subject, predicate, object). This language is widely used in the semantic world. We use an well-known open source solution called Jena<sup>8</sup>.

In this solution, an ontology server works alongside the workflow of data (referred to as the dual model).

---

<sup>8</sup> <http://jena.apache.org>

## The dual model



The models are defined in the Meriterm Online Tool (MOT) and stored in an object oriented database (object abstraction provided by a middleware providing persistent storage of objects). Then the models are processed by tools to create files to be used by Jena ( a JAVA API for RDF- to create or update models stored in TDB (RDF storage). Fuseki is a SPARQL server to serve RDF data over HTTP.

This approach permits to store semantic and non-semantic models : ICPC, ICD, ... and TMF, LMF, Lemon, ... and make links between these.

An easy to use human web interface is critical to manage the models and make links between them. Non-semantic models are generally 2D (easily stored in a spreadsheet). Semantic models are more complex (stored as triplets, with multiple links to other triplets). The user interface will provide a way to visualise all these links and see the data from different angles.

## A web-based application

The tool will be available online and build using internet standards like HTML5 and Java-script.

The languages used to build the backend will be defined regarding the task and interfaces needed by the standards and the other software used.

## Final remarks

As experienced in many projects trying to build a similar resource (mapping ICPC, ICD, Locas, Snomed), the quality tool is the key to the success of the project.

Without a dedicated tool, the process is painful and only very motivated people are willing to make it to the end.

The quality of the resource is also strongly related to the tool used when building it.

For a project to achieve its goal of building a quality resource and maintaining it (release after release), a well-thought tool is essential.