Build-up of Terminology Services for the European Centre for Disease Prevention and Control (ECDC) – part two: conceptual engineering

Gergely Héja, László Balkányi

Abstract

This paper is the second part of the adjoined papers describing the chain of thoughts leading to plan and implement a terminology system at ECDC – describing the conceptual model of the ECDC terminology system, which is built on a formal ontology backbone. ECDC – an EU scientific agency with the scope of preventing and controlling communicable disease – is an organisation started in 2005. In it's short history it has initiated the build-up of a number of information systems – services in parallel. In part one, reasoning was given for a common, shared terminology service, providing proper interface both for human and machine users. Such a service seemed to be the proper tool ensuring a consistent conceptual space among the different systems, notably ensuring transparency and cross-search ability. This paper explains the conceptual modelling in the form of a formal ontology as the conceptual backbone for so called value sets representing terminology sources. The value sets are modelled in SKOS.

Keywords: formal ontology, biomedical ontology, OWL, terminology services, public health, SKOS

Introduction – backgrounds

Following the chain and the results of thoughts leading to plan a terminology system at ECDC a conceptual engineering pathway was envisaged. This paper explains the devised conceptual model, the standards chosen and the ways how different conceptual sources are used and aligned to achieve transparency and cross search abilities.

The heterogeneous information systems at ECDC [1] (see paper one) require a common terminology service with a shared conceptual system. These information systems make use of various conceptual systems ranging from plain lists of terms through taxonomies to complex thesauri, which have to be managed in a common terminology. The main conceptual sources of this terminology are:

- database tables (values for entry forms) of TESSy (The European Surveillance System) and TTT (Threat Tracking Tool)
- keyword list of Eurosurveillance and the Influenza Horizontal Project
- terms extracted from legal and scientific documents related to funding of ECDC and activity of scientific staff
- (parts of) "external" terminologies such as ICD10 [2] or SNOMED CT [3]

Methods

– the logical conceptual model

The conceptual sources are represented as *value sets*, the terms as *categories* in the particular value set. The categories may be interconnected by hierarchical and other relations (depending on the conceptual structure of the source). Each category is mapped to exactly one *concept* in the common *ontology* specifying the meaning of the particular category. Categories from different value sets may be mapped to the same concept thereby automatically interconnecting these categories. There may be several hundred thousand categories (e.g. if ECDC integrates large parts of external reference value sets such as SNOMED CT in the conceptual system or the lists of bacteria serovars), but the formal reasoning on this scale is troublesome, consequently the number of concepts in the ontology should be kept relatively low (in the order of ten thousand). For that reason we allow that the meaning of the concept is broader than the mapped category. Categories may have natural language labels: a preferred term (in a given language) and synonyms. We use Dublin Core [4] to add annotations to the value sets and categories.

Certain value sets (e.g. that representing ICD10) change over time, consequently the terminology system has to support history tracking, implemented by versioning of the categories. For easier implementation we decided that a version contains the whole information (labels, description, relations and ontology binding) pertaining to the category in the given time interval, instead of storing only the difference. It is also possible that categories are deleted, split or merged.

The ontology is an extension of EDnS (Extended Descriptions and Situations) [5], described in OWL DL [6]. EDnS is a module extending the formal top-level ontology DOLCE [7] with concepts and properties representing descriptions and situations. For easier knowledge management and maintenance we decided to create a modularised ontology with modules describing:

- "High-level" concepts representing anatomy, with concepts derived from the FMA (Foundational Model of Anatomy) [8]. Since the FMA is an huge reference ontology with approximately 70 thousand concepts, we took only those needed for the purposes of ECDC: organ systems, organs, and some other entities (like blood or joint). Anatomical entities are modelled either as edns:non-agentive-physical-object (e.g. organs) or dol:amount-of-matter (e.g. tissue or body substances).
- (Pathogenic) organisms, both living (Bacteria and Eucarya) and non-living (Virus), classified according to biological taxonomy. Living organisms are modelled as edns:agentive-physical-object. We decided to classify all living organisms as agentive, because in DOLCE Intentionality is understood "… as the capability of heading for/dealing with objects or states of the world.", and bacteria have this capability, while viruses not.
- Chemicals and biological (macro)molecules are modelled as edns:non-agentive-physical-object, while toxins, antibodies, drugs etc. are modelled as edns:role played by the corresponding molecule. Drug roles are classified according to ATC.
- Pathology, modelled typically as a process-structure pair (like in GALEN [9]), e.g. inflammation process (dol:accomplishment) and inflammation lesion (dol:feature). We decided to represent processes as dol:accomplishment instead of dol:process (as suggested by the documentation of DOLCE) because various implementations (e.g. EDnS and OWN [10]) classify processes

as dol:accomplishment..

- Diseases, are modelled as edns:description, with three main characteristics (if possible):
  - Pathology (e.g. inflammation)
  - Main anatomical location (e.g. respiratory tract)
  - Etiology (e.g. Legionella pneumophila)
- (Public) health activities, agencies, roles:
  - (Public) health activities (e.g. surveillance, vaccination) are classified as an edns:activity (since they are planned).
  - (Public) health events (e.g. outbreak) are classified as an edns:accomplishment (since they are typically not planned).
  - Laboratory, etc. methods and procedures are classified as an edns:method and edns:activity.
  - Index patient, primary laboratory, etc. are classified as an edns:role.
  - Agencies (e.g. ECDC, WHO) are classified as an soc:organisation.
  - Legal concepts are classified according to the DOLCE extension CoreLegal [11].

The devised ontology is mainly a conceptual core modelling ECDC-relevant concepts on different granularity levels (according the currently available conceptual sources). It is expected that the ontology will be significantly extended in the near future.
The ontology contains only classes, and no individuals, consequently an A-BOX [12] capable formal reasoner is not necessary, T-BOX suffices (e.g. FACT++ [13]).
The content of the ontology may change over time, e.g. the biological taxonomy changes, species are reclassified. To represent this change in knowledge, the concepts have three annotation properties:

- An optional obsolete flag, denoting whether the concept is obsolete
- precededBy (cardinality 0..*, because concepts may be merged) identifying the concept from which the actual one is created
- succeededBy (cardinality 0..*, because concepts may be split) identifying the concepts which are created from the actual one

Although the ontology, organised according to the structure enforced by DOLCE enables conceptual consistency across the relevant domains, it might be quite confusing to the typical user of the ECDC terminology system: the epidemiologist. Therefore the ontology will be converted to a semantic network 'hiding' the formal logical definitions of the concepts, only showing the (for-the-epidemiologist) relevant hierarchical and non-hierarchical relations between them. Moreover the concepts related to a certain (sub)domain (e.g. public health) are placed under several remote concepts in DOLCE. Consequently top-level domain concepts (e.g. public health) are needed in the semantic network to help the navigation. These concepts will be generated during the transformation of the ontology from logical representation to the semantic network.
This transformed ontology is represented with the classes

- *Ontology*, representing the ontology
- *Concept*, representing the classes in the OWL files
- *Relation*, representing the definition of object properties in the OWL file

The logical model of the value sets and the semantic network can be seen in Figure 1. The class *Description* contains the Dublin Core annotations used by SKOS.

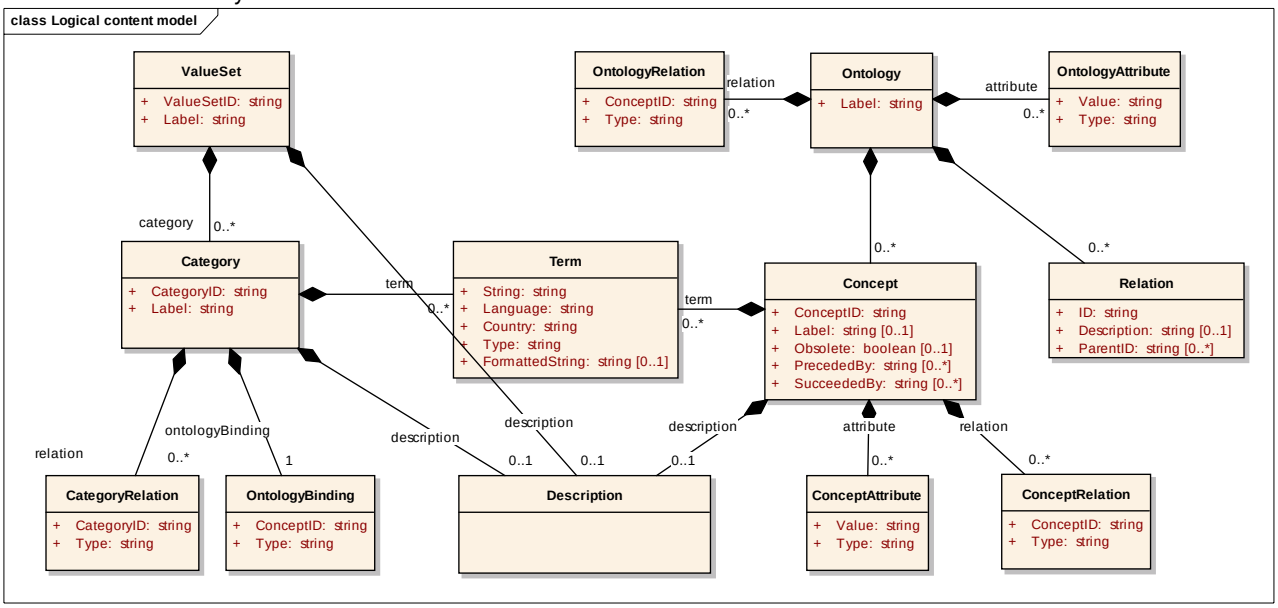

Figure 1 – Logical model of the terminology system

– the implementation of the conceptual model

We decided to use SKOS [14] and SKOSMAP [15] to represent the content of the terminology system. The value sets and the ontology is represented as skos:ConceptScheme, while categories, concepts and relations are represented as skos:Concept. The relations between categories are represented as sub-properties of skos:semanticRelation. The mapping of categories to concepts are represented as sub-properties of skosmap:mappingRelation (only exactMatch and broadMatch is allowed). The relations between concepts are represented as sub-properties of skos:semanticRelation. The implementation is represented in Figure 2.

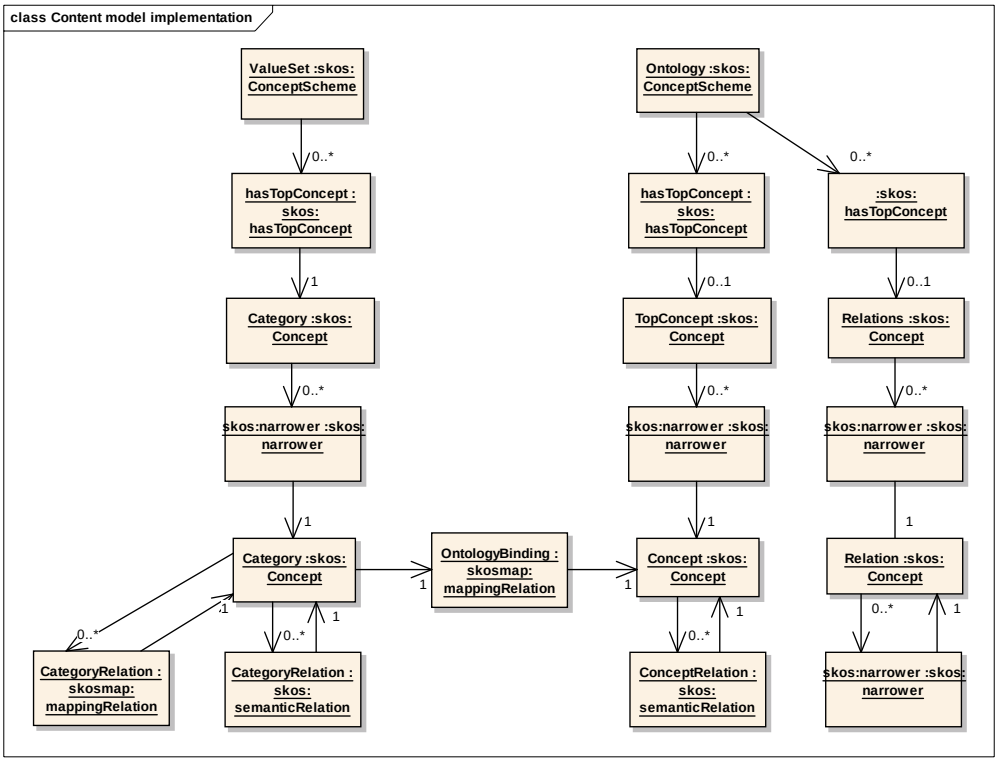

Figure 2 – SKOS-based implementation of the logical model

Discussion and conclusions

Regarding the three main areas mentioned in part 1, i.e. the need for a common conceptual standard, the need to ensure long term stability and the need to handle multilingualism, use-specific formatting and other labelling variations of concepts, the experiences of modelling work are:

- From the tried conceptual structures SKOS seems to be the most appropriate format to represent heterogeneous terminological sources.
- ECDC activities require the handling and conceptual interoperability of heterogeneous conceptual systems. Consequently we decided to use an ontology to create the common conceptual framework. It also provides the paradigm for future expansions. The formal definitions enable formal consistency checking of the conceptual system, moreover it will also provide intelligent querying and navigating services by automatic classification.
- ECDC operates in a multilingual setting. The concepts and categories provide the language-independent conceptual framework, while the multilingual natural language labels add the needed understandability for human users across the EU. Certain users of the system require the representation of formatted terms (e.g. for report generation). This need has been solved by using an external collection of such formatted terms, referred to from the SKOS description.

Since the terminology service is only in planning phase, we do not have practical results of the conceptual system at the time being. The fine tuning of the semantic network for user friendliness can only be achieved after the pilot operation of the terminology server envisaged for the end of 2007.

References

[1] www.ecdc.europa.eu
[2] International Statistical Classification of Diseases and Health Related Problems, Who, Geneva, 1992
[3] http://www.snomed.org/snomedct/index.html
[4] http://dublincore.org/
[5] http://wiki.loa-cnr.it/index.php/LoaWiki:DnS
[6] http://www.w3.org/2004/OWL/
[7] http://www.loa-cnr.it/DOLCE.html
[8] http://sig.biostr.washington.edu/projects/fm/
[9] http://www.opengalen.org/open/crm/index.html
[10] http://www.loa-cnr.it/DOLCE.html#OntoWordNet
[11] http://wiki.loa-cnr.it/index.php/LoaWiki:CLO
[12] F. Bader et al. (editors): The Description Logic Handbook (Theory, Implementation and Applications), Cambridge University Press, 2003
[13] http://owl.man.ac.uk/factplusplus/
[14] http://www.w3.org/2004/02/skos/
[15] http://www.w3.org/2004/02/skos/mapping/spec/

Correspondance:

Gergely Héja M.Sc.., knowledge engineer, e-mail: heja@mit.bme.hu
Falcon Informatics Ltd, 27. Násznagy str., Budapest, Hungary, H-1131