Build-up of Terminology Services for the European Centre for Disease Prevention and Control (ECDC) – part one: enterprise- and functional-level planning

László Balkányi[1], Gergely Héja[2]

Abstract

ECDC (an EU scientific agency with the scope of preventing and controlling communicable diseases) is an organization started in 2005. ECDC has initiated the build-up of a number of domain specific information systems in parallel. A common, shared terminology service, providing proper interface both for human and machine clients seemed to be the proper tool towards conceptual consistency (later interoperability) among the different systems, notably ensuring transparency and cross-search ability in the quickly accumulating communicable diseases related data, information and knowledge. There will be internal and external machine clients of the terminology service, as the new European communicable disease surveillance system (TESSy), the epidemics intelligence supporting information systems (EPIS, TTT[3]), the ECDC WEB site, the Eurosurveillance journal, the Knowledge and Information service (KISatECDC[4]) etc. The human users will be internal and external experts, the public health and the communicable disease scientific community. This paper describes the enterprise and functional level planning of the terminology services, the necessary decisions ECDC were faced at and the planned process to build up services. As it was found that there will be a need for long term, strictly maintained, conceptually clear terminology handling to ensure consistency and expendability, the build-up of an ontology backbone became a core part of planning. The technical and engineering aspects are given in an adjoining paper.

Keywords: terminology services, communicable disease, information systems, planning, biomedical ontology

Introduction – backgrounds

The paper describes the chain of thoughts leading to plan and implement terminology services at ECDC – up to the phase of enterprise and functional level planning, the standards chosen, and the ways how term sources are used and aligned to achieve transparency and cross search abilities – with other words how to use the planned term services as a semantic glue among different information services of a quickly developing organisation. A second, joint publication discusses technical, engineering aspects leading to specification of services.

ECDC is an EU agency that has been created to help strengthen Europe's defense against infectious diseases. The mission of ECDC is to identify, assess and communicate current and emerging threats to human health posed by infectious diseases. ECDC works in close partnership with member state health protection bodies across Europe. Among others, the aim is to pool Europe's health knowledge in this field and to develop authoritative scientific opinions about the risks posed by current and emerging infectious diseases. The Centre, currently staffed with ~ 80 experts and management, will have an over € 50 million budget by 2010 and its staff will grow to 300 over the coming years. ECDC has a matrix organizational structure with four (vertical) technical units (Scientific Advice, Surveillance, Preparedness and Response, Health Communication), an administrative unit and seven horizontal disease-specific projects (like for Tuberculosis, HIV-AIDS, Influenza, etc.) All units develop & operate information systems as major tools to carry out their tasks. These systems include: (a) TESSy – the new, unified European communicable disease surveillance system that will comprise all previously independently managed disease specific surveillance networks and the basic surveillance networks, (b) EPIS, the epidemics intelligence portal that will support event detection / risk assessment / outbreak investigation / control measurers on EU level, (c) the ECDC WEB site, (d) the Eurosurveillance journal, (e) the Knowledge and Information service (KISatECDC), that will be the content managing system for scientific documents in ECDC. While all these information systems use the shared physical environment and back-office services of the ECDC information system, it is quite clear that besides the shared 'bitways' there has to be a shared semantic space for those systems - although the systems cover quite different contextual domains. While in an ideal situation the same concepts should have the same labels across all these services, in reality this is not possible due to specific professional reasons. Therefore a common reference terminology has to be established, mapping the different 'use-specific' labels onto a common conceptual scheme. If ECDC lacks this, there will be no transparency, no cross checking, no cross searching among the information islands of the mushrooming separate systems, also significant resources will be wasted due to parallel work on terminologies maintained for each of the systems independently. An additional strong argument for a common (and also structured!) semantic space is the chance for navigation support by adding relations, adding structure to simple lists of concepts – which is implemented as an 'ontology backbone'. At least the following domains should be brought under a common conceptual umbrella: epidemiology, public health, clinical medicine, 'Euro legalese', EU member states legalese, WHO, etc. Although there is a significant overlap across these domains, especially in the medical science area, still the differences of meaning and granularity are far from being negligible. The consequences of historical changes of concepts / labels has to be solved as well, as the timeline related to the 'semantic space' of the agency covers many decades – backwards and forwards.

---

[1] ECDC staff member

[2] contracted expert to ECDC

[3] EPIS: the to-be-developed epidemic intelligence portal, TTT: the operational Threat Tracking Tool

[4] KISatECDC: internal electronic content management with windows to the external scientific community

This situation leads us to the notion of building terminology services, with the basic task of serving these applications as well as the human users, keeping semantic consistency, at least navigability across the systems and the domains.

Methods – tools

ECDC had to come up with a set of interlinked methods and tools resulting in the build-up of the terminology services. Here we will give the principles of choosing the tools /methods and a list of them. In the discussion part we will come back to some alternatives and issues related to the methodology. The following general principles guided the choice of methods: avoiding reinventing the wheel; using standards as far as possible, building consensus with other system developers/users; compliance with existing ECDC software and hardware environment; spending public money with utmost care. All tools and methods should enable / allow multilingualism and support / allow a time machine functionality that is to handle changing versions of concepts and their labels along the time line of several decades. Keeping this principles costs time, but hopefully this will result in enduring, robust services. We decided to take four steps: (a) *an analytical step* to fully understand and map sources of concepts, (b) *extraction of* concepts and related labels, (c) *setting up a core terminology* using the extracted material; *(d) conceptual design and specification of terminology services* on the enterprise, on the functional and on the logical levels. This paper deals with enterprise and the functional level, while part two deals with the logical level issues. The following tools and standard methods were used: command line tools and disk cataloguing utilities for extraction of file/folder structures (as file/folder names will be used as metadata, as keywords to the documents stored), a Unicode text editor for building/editing xml files, a spreadsheet application for file statistics and term handling for human users, GATE [1] for term extraction from texts, Protégé [2] for ontology editing, SKOS [3] for building a common structure for differently structured external and internal value sets (like e.g. International Classification of Diseases [4]), DCMI [5] for structuring object metadata, OWL [6] for the ontology, a UML [7] tool to build use cases, class diagrams and component model of terminology services

Results and planned further steps

For step (a) the result of understanding the set-up of concept sources in ECDC is crystallized in a content map, that will allow ECDC users to navigate in the content space by the general query of 'Who (actors) does (activities) what (topics)?'. For step (b) the results are sets of terms from the very differing internal and external sources like official ECDC legal and scientific documents [8], planned TESSy and TTT database variables, various ECDC spreadsheet field names, internal lists and relevant external classifications, thesauri, nomenclatures etc, called together 'value sets'. To distinguish the individual items in these very differing value sets from the final terms in the terminology services, we call these items 'categories'. The resulting category sets can be grouped in three main groups: simple lists of terms (example: keyword list of Eurosurveillance [9]), tree hierarchies (taxonomies, an example is ICD 10, although it is not just a hierarchy) and complex nets of terms (e.g. MeSH [10] and SNOMED CT [11]). *Step (c), setting up the core terminology* was two activities in parallel: (1) To find and apply a method, that is able to handle all the three kinds of sets with a common, standard data model. Here the final decision was to use SKOS that is an upcoming W3C standard based on RDF. The product here is a series of SKOS (RDF) files. (2) At the same time, based on experience and literature [12] we have realised the need for an ontology backbone to all these value sets – this will be explained in the Discussion section below. To build this ontology we have used Protégé. The product of this activity is a collection of OWL files that contains the core ontology, defining the concepts 'behind' the collected terms of the value sets. The categories of value sets are mapped to the concepts in the ontology by SKOSMAP. The details of conceptual modelling are explained in part two. For *step (d)* the result is a technical specification document containing textual description and UML diagrams of the terminology services, that will serve the ECDC (and external) applications as well as human users with terminology services. This is the basis of an ongoing development work – further steps will be implementing the terminology server / services than to extend the ECDC core terminology and make it available for the scientific community as a resource via the Web. The OWL DL modelling of the ontology enables us to plan for more value added functions that will use automated reasoning in the future. The current task is however limited to the terminology services directly serving the application needs.

Discussion

Three main areas might be highlighted among the numerous problematic issues that had to be dealt with:
First was to find a common standard that is capable to model all the differing conceptual structures of the sources. For this the SKOS formalism seems to be an adequate answer. SKOS "provides a model for expressing the basic structure and content of concept schemes such as thesauri, classification schemes, subject heading lists, taxonomies, ,.... also concept schemes embedded in glossaries and terminologies" [3], being "an application of the Resource Description Framework (RDF), that can be used to express a concept scheme as an RDF graph". SKOS features as explained in part two give us the needed flexibility. An additional chosen format for value set export and import (CLaML,[13]) enables term exchange with WHO health classifications.
Even if SKOS allows term and terms relation modelling saving the 'structural information' what we need from the external and internal sources, we still need a more robust level of concept representation that is independent of actual values of value sets. For this reason and to ensure advanced knowledge management functions to be developed later, we thought that there is need to find a solution to our *second problem* that is to ensure long term stability, reusability and extendibility at the same time in our terminology services. For these reasons we have to have a stable ontology behind the (changing) SKOS formatted concept sets, where the changes are in the hands of the external publisher of the value set, e.g. the change from ICD 10 to ICD 11 in the future. The ECDC core domain ontology will remain stable as long as our domain knowledge will not change – while the value sets will be refreshed according their versioning. (I.e. it is very probable, that the lung will remain part of the respiratory system,

and the pneumococci will have causal relation to the pneumonia, that itself will remain a disease of the lungs.) There was not much debate on using OWL for this purpose. Details are given in part two.
A *third issue* was how to handle the multiple formalisms of value sets (including the multilingualism and also the special formatting needs by some of the user applications). We have decided to have a conceptual 'three box' approach in handling the terms. The first conceptual box contains the ontology. The second conceptual box contains a number of SKOS modeled value sets. A typical machine user, that uses the terminology server as its terms source, will use only its own value set - while all the maintenance issues of this value set are handled by the terminology services administration. Besides this single value set utilization some future machine users (like the Epidemics Intelligence Portal) will have to gather information from more than one systems as the TESSY and the TTT – for this query it will have to get back all the relevant categories of the value sets of these applications, using the ontology bindings. In that case a third conceptual box is a generated sum of terms from the queried systems. The human user will have also the functionality of querying one value set or conceptual 'intersections' or 'unions' of more value sets, thereby 'generating' a query specific third box of terms. Value sets and the ontology are edited only by the terminology service administrator. The generated output can handle the multilingualism and the use-specific formatting utilizing (in-SKOS-model-embedded) knowledge. This approach also enables a two level 'time machine', differentiating between the simple version changes in a value set and the change in the scientific knowledge that is represented in the ontology.
Summing up, there will be three ways to 'interact' with the terminology server: (1) importing an ontology (edited externally i.e. by Protégé), (2) importing, exporting and editing value sets; (3) making use-specific and general queries for generated terms and their relations. The first two interactions are available for the terminology system administrators, the third, complex interaction is aimed at the end-user, that can be human or a software actor.

Conclusions

We conclude that the above described project results shows us that the state of the art in applied terminology – ontology science is adequate to handle a very complex situation; it is possible to align terms from differing sources under a common conceptual 'umbrella' by reusing models of the literature. Standards and tools available on the Web enable developers to set up well defined, conceptually consistent term structures for software developers.
We also conclude that even if in a given professional domain there is no ready made solution for a unique domain specific terminology – like in our case for epidemiologic activity for communicable diseases, there is no need to start from scratch as (usually) it is thought to be the only way to guarantee consistency and specific usability. The wide variety of available broader context domain sources enable us an also press for re-utilization - thereby greatly reducing the effort leading to establishment of usable terminology services.

References

[1] http://gate.ac.uk/
[2] http://protege.stanford.edu/
[3] http://www.w3.org/2004/02/skos/
[4] http://www.who.int/classifications/icd/en/
[5] http://dublincore.org/
[6] http://www.w3.org/TR/owl-features/
[7] http://www.uml.org/
[8] Regulation (EC) no 851/2004 of the European Parliament and of the Council of 21 April 2004 establishing a European Centre for Disease Prevention and Control; Decision No 2119/98/EC of the European Parliament and of the Council of 24 September 1998 setting up a network for the epidemiological surveillance and control of communicable diseases in the Community; 2000/96/EC: Commission Decision of 22 December 1999 on the communicable diseases to be progressively covered by the Community network under Decision No 2119/98/EC of the European Parliament and of the Council; 2003/534/EC: Commission Decision of 17 July 2003 amending Decision No 2119/98/EC of the European Parliament and of the Council and Decision 2000/96/EC as regards communicable diseases listed in those decisions and amending Decision 2002/253/EC as regards the case definitions for communicable diseases; 2003/542/EC: Commission Decision of 17 July 2003 amending Decision 2000/96/EC as regards the operation of dedicated surveillance networks
[9] http://www.eurosurveillance.org/
[10] http://www.nlm.nih.gov/mesh/
[11] http://www.snomed.org/snomedct/index.html
[12] Ontology and Medical Terminology: Why Description Logics Are Not Enough, Werner Ceusters, Barry Smith, Jim Flanagan, at: Towards an Electronic Patient Record (TEPR 2003), San Antonio 10-14 May 2003, Boston, MA: Medical Records Institute
[13] CEN TC251 Technical Specification (TS14463).

Correspondence:

László Balkányi MD. PhD., knowledge manager
European Centre for Disease Prevention and Control (ECDC)
Tomtebodavägen 11A, SE-171 83 Stockholm, Sweden
E-mail: Laszlo.Balkanyi@ecdc.europa.eu
Web: http://www.ecdc.europa.eu