Building and Using Terminology Services for the European Centre for Disease Prevention and Control

László Balkányi[1], Gergely Héja[2], and Cecilia Silva Perucha[1]

[1] European Centre for Disease Prevention and Control,
Tomtebodavägen 11a, Stockholm, Sweden
[2] Falcon Informatics Ltd., Szentlászlói út 189, Szentendre, Hungary
laszlo.balkanyi@ecdc.europa.eu

Abstract. This paper describes the process of building terminology service and using domain ontology as its conceptual backbone for a European Union agency. ECDC, established in 2005, aims at strengthening Europe's defences against infectious diseases, operates a range of information services at the crossroads of different professional domains as e.g. infectious diseases, EU regulation in public health, etc. A domain ontology based vocabulary service and a tool to disseminate its content (a terminology server) was designed and implemented to ensure semantic interoperability among different information system components. Design considerations, standard selection (SKOS, OWL) choosing external references (MeSH, ICD10, SNOMED) and the services offered on the human and machine user interface are presented and lessons learned are explained.

Keywords: ECDC, domain ontology, terminology services, semantic interoperability.

1 Introduction

1.1 Understanding the Problem – Why ECDC Needs Interoperability Tools?

The European Centre for Disease Prevention and Control (ECDC) builds and operates a range of information services at the crossroads of different professional domains as public health, microbiology, EU regulations. Examples are: The European Surveillance System (TESSy), Threat Tracking Tool (TTT), Intranet Document Repository (content services), Expert Directory. As most of ECDC systems are 'one of their kind' and they were developed under a time pressure to become functional as soon as possible, it was inevitable that on layers of logical design, data modeling, content phrasing, user interfacing, etc., the systems started to diverge. From early on ICT services in ECDC, under sever operational strain, achieved a certain homogeneity on the 'bit-ways' level of operating systems, networking, communication and for the mundane office tasks (served by off-the-shelf products) by choosing tools of one software provider. On the other hand, in between the top, presentation layer and bottom, bit-ways layer, in the middle layer of shared services and contents neither the planning nor the design and implementation of systems have been aligned. Following a careful

P. Kostkova (Ed.): eHealth 2009, LNICST 27, pp. 116–123, 2010.

analysis of the emerging discrepancies and multiplication of coupling needs of different systems, a solution was offered to set up respective layers of interoperability tools / methods / project alignment measures that will on the long run ensure seamless flow of information. This paper focuses on a tool aimed at solving the problems of one layer, the semantic interoperability.

1.2 Semantic Interoperability and the Selected Tools: Building a Terminology Server and a Domain Ontology

It is not a surprise, that complex, knowledge intensive organizations, like ECDC working at crossroads of different professions will suffer from the inconsistent labeling of the same concepts. A long history of using enterprise wide data vocabularies in medicine [1], even (in a way) the emergence of markup languages [2] themselves are all symptoms of this phenomenon. The ICT tower of Babel happens to happen over and over again in growing organizations. Analyzing the situation at ECDC [3] it became clear, that implementing simply a rigorous data modeling rule set plus introducing obligatory terminology is not an option because of several reasons: (1) Highly trained professional users (at different, although closely related domains) are prone to use their own professional jargon. (2) Users need different 'granularity', different level of precision in different settings. (3) Grouping, typing, classifying concepts are especially prone to the very specific needs of the given function – you can group communicable diseases according their etiology, the symptomatology of the caused disease, the needed public health measures, the vectors, etc. Therefore ECDC needs a very flexible solution that on one hand allows the specialist to use their known terminology in their known context, on the other hand will gently guide these diverse groups toward a unified terminology, allowing different depth of granularity. To achieve these goals two years ago it was decided that a centrally administered terminology service will be provided, a software application will be built. An agency specific ontology should become the conceptual backbone to connect, to cross map the already existing particular term sets and pushing system users, developers toward a more homogeneous, consistent, navigable multi-domain terminology.

2 Methods, Tools, Standards

2.1 Avoiding Reinventing the Wheel – What Is Out There?

In order to avoid reinventing the wheel, a survey was done focusing on similar organizations.

Terminologies, classifications: Among the (literally) several hundreds of existing medical terminologies we mention here only a subset of them, with very different scope and origin. All of them are relevant as sources for a multi-domain terminology in the area of communicable disease prevention and control, in public health:

The CDC VADS (Public Health Information Network Vocabulary Access and Distribution System [4]) is itself a multitude of vocabularies and a service which allows public health partners to browse, search and download concepts. VADS is huge, with a very rich and broad scope of value sets. Concepts inherit the semantics from the

coding system associated with the PHIN vocabulary domain in which they are placed – but there is no overarching semantic model behind the value sets.

The WHO is mandated to the production of international classifications on health. The purpose of the WHO family of classifications (WHO-FIC) is to promote appropriate selection of terms for health fields across the world. It consists of the (1) International Classification of Diseases (ICD), (2) the International Classification of Functioning, Disability and Health (ICF) and (3) the International Classification of Health Interventions (ICHI). All of them have a long legacy compromise along different points of views, and are sometimes abused. Their conceptual frames are not being based on information science principles but on health statistics pragmatics [5].

The largest medical knowledge 'body' made available on the WWW is Medline of the US National Library of Medicine. One of its resources is MeSH (Medical Subject Headings), a controlled vocabulary thesaurus. It consists of sets of terms, descriptors in a hierarchical structure that permits searching at various levels of specificity. There are 25,186 descriptors in MeSH 2009. It is very robust, very well documented and supported collection, but the same way as all the previously mentioned conceptual systems, it lacks a scientifically proved semantic model, enabling e.g. machine inference, or 'logical calculability' [6].

A remarkable recent effort overcoming this lack of being not based on the principles of information science and having relevant, reusable medical or health related content at the same time, resulted in a multi domain health ontology development in the medical arena, called the OBO foundry. This is a 'collaborative' experiment. It involves developers of science-based, health or medicine related ontologies. The developers agreed on a set of common principles for ontology development. The goal is creating a suite of orthogonal interoperable reference ontologies in the biomedical domain [7].

Regarding **standards,** CEN TC 251 and ISO TC 215 are the Technical Committees of standardization in medical informatics on the European and World level [8]. Limitations of current ISO and CEN standards on medical terminology are explained here [9]. International Health Terminology Standards Development Organization is the current steward of Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) [10]. SNOMED CT is considered to be the most comprehensive, multilingual clinical healthcare terminology in the world.

Health Level Seven (HL7, [11]) is a US accredited Standards Developing Organization (SDO) operating in the healthcare arena. HL7's domain is clinical and administrative data, it develops specifications. It is probably the most widely used messaging standard that enables disparate healthcare applications to exchange key sets of clinical and administrative data. It has to be mentioned, that there are several serious problem areas, like documentation (intelligibility), implementation, quality of internal consistency of the HL7 RIM etc [12].

Obviously medicine is not the only field where multiple disciplines meet and where huge amount of differently organized and classified information has to be efficiently stored, merged, searched, etc. The Worldwide Web itself is the best source for checking these challenges. Widely used and emerging W3C standards (described later in more details) allow a (hopefully) future proof information modeling and storage in data formats that can be used by a wide array of applications, thereby ensuring transparency.

2.2 What Methods, Tools and Standards Have Been Chosen?

To ensure proper support and engineering level integration in ECDC, the terminology server application (TS) was built on MS platform, using an MS SQL engine. Nevertheless for the machine level communication the chosen method was to communicate via standard web services using SOAP interface to allow other applications based on any platform to communicate with the TS. The current core TS uses a predefined set of atomic queries, that can be combined, but the next version already under development will allow SPARQL [13] queries.

Terminologies are represented according to the format of an emerging W3C standard, SKOS [14]. SKOS provides a standardized way to represent knowledge organization systems (as e.g. thesauri, taxonomies, classifications and subject heading systems) using (a limited subset of) OWL. SKOS allows information to be passed between applications in an interoperable way. SKOS also allows knowledge organization systems to be used in distributed, decentralized metadata applications.

The ontology, that serves as conceptual backbone and where (other than just hierarchic) relations among concepts are stored, is stored as an Ontology Web Language (OWL) file. The decision to represent the ontology in OWL rather than SKOS is due to some limitations of SKOS in expressing relations and allowing certain operations among terms. OWL, another W3C standard, was designed for use by applications that need to process the content of information instead of just presenting. OWL can be used not only to explicitly represent the meaning of terms in vocabularies but also all kind of relationships between those terms [15]. OWL has more facilities for expressing meaning and semantics than SKOS, and thus OWL goes beyond in its ability to represent machine interpretable content on the Web.

3 Results in Building Terminology, Operating the Terminology Server (TS) and Planned Next Steps

Content: The TS contains two `content levels`: `value sets` on the bottom level and `ontology` on the top level. On the *`value set` level* there are three collections of terminologies: (1) application specific sets, like variables of the Threat Tracking Tool, (2) common, shared value sets, e.g. Pathogen Organisms and (3) external reference value sets, e.g. ECDC relevant MeSH terms.

For each term we store a preferred label, alternative labels, a valid/obsolete flag, as well as other metadata items. If a term becomes obsolete, there is a relation to its succeeding, new term. Relationships to `parents` and `children` are stored as well - plus a binding relation to a concept in the backbone ontology. This binding relation ensures a `mapping` of a term`s meaning from one value set across the ontology to any other value set.

On the ontology level ECDC ontology uses DOLCE [16], a domain neutral top level ontology for top level concepts, and tries to build scientifically correct domain ontology. The ontology currently describes concepts the following domains:

- organism, based on the biological taxonomy. Most organisms are human pathogens, but hosts (e.g. swine) and vectors (e.g. mosquito) are also listed.
- anatomical structures, currently on organ and organ system level. The anatomical model is based on FMA [17].

Table 1. Existing and 'under construction' value sets in ECDC Terminology Server

Terminology set types	Examples	Estimated size (no. of terms)
Application specific sets:	Threat Tracking Tool (TTT)	~ 160
	The European Surveillance System (TESSY)	~ 280
	Web Portal topics	~ 85
Common, shared sets:	Pathogen Organisms	~ 4700
	Communicable diseases, conditions, syndromes	~ 1200
	Public health terms	~ 1600
	Geo entities (countries, cities, regions)	~ 40500
	Organizations	~ 400
	Administrative terms, abbreviations, acronyms	~ 200
External reference sets:	ICD 10 (relevant subset)	~ 1200
	MeSH (relevant subset)	~ 3760
	SNOMED (relevant subset)	~ 16500
	Sum:	~ 70600

- diseases, in the mandate of ECDC are represented, however the extension to all infectious diseases from ICD10 is under way.
- medical and epidemiological actors, activities and events.

The ontology is intended to be used later for automatic reasoning - consequently it is represented in OWL DL. The tool used for ontology building is Protégé. Because the top-level of the ontology by its nature deals with quite abstract notions, in its original from is not very usable for physicians. For that reason during the upload of the ontology into the TS it is converted to a semantic network that is easier to understand by physicians. Table 2 displays the top-level of this (simplified) semantic network.

Table 2. ECDC domain ontology, top level classes and properties

Generic classes
Activity
Anatomy
Biological process
Collection
Disorder
Equipment
Geopolitical entity
Legal regulation
Material
Organization
Organism
Personal role

Properties		
attribute		
relation		
	caused by	causes
	has member	member of
	has temporal feature	temporal feature of
	locative relation	inverse locative relation
	partitive relation	inverse partitive relation
	produced by	produces
	resistance of	resistant to

Functions of TS: The Terminology Server answers queries, using a basic set of them that can be combined for more sophisticated querying. The client applications use this set of queries retrieving terminology information (synonyms, related terms, time line of changes, etc). The terminology server has also a human user interface, allowing experts to browse all the value sets, and the ontology, to navigate in this term 'space' along the defined relations among categories (of value sets) and concepts (of ontology).

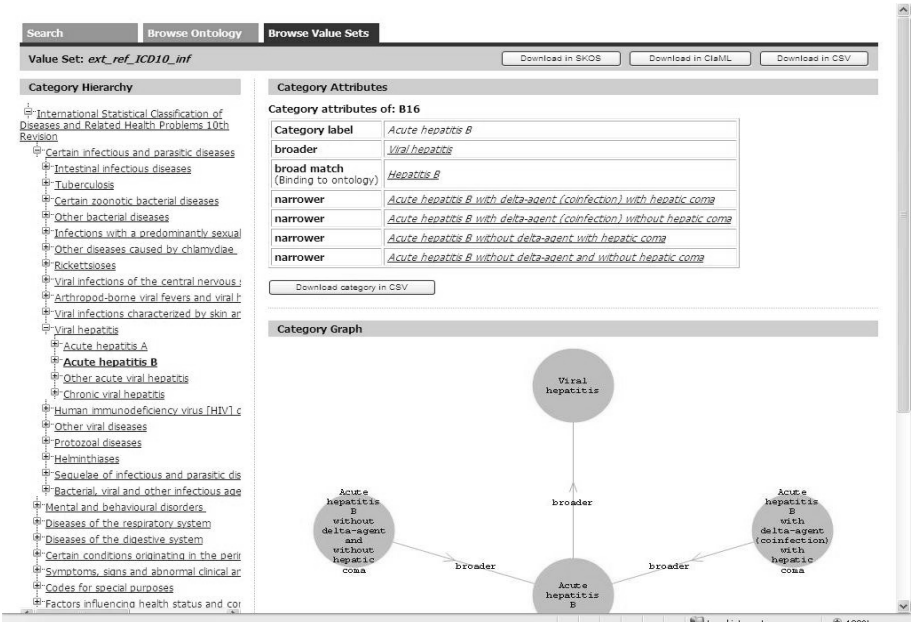

Fig. 1. Screen shot of human user interface of ECDC core terminology server

The extended version that is under development at the publication of this paper will provide enhanced complex query services, query using SPARQL expressions, alert notification to value set administrators, etc. The future plans include Web publication of ECDC terminology services, made available for public health system developers of EU member states, both on the human and the machine interface.

Table 3. Services of the core terminology server

Functionality	Explanations
Read operations:	*Domain of information:* • Ontology, concepts and relations; • Value sets and categories. *Operations:* • Get detailed or short description of data elements; • Navigate the data elements according to the retrieved relations; • Search of data elements; • Download the ontology and value sets in OWL, SKOS, ClaML, ... format.
Write operations:	*Domain of information:* • Ontology, concepts and relations; • Value sets and categories. *Operations:* • Importing the ontology and value sets (SKOs, OWL) • Creating, modifying and deleting value sets and categories.
Retrieval of a certain conceptual element:	E.g.: GetValueSets, GetValueSet, GetCategory, GetOntology, GetConcept and GetRelation);
Search for conceptual elements:	With matching the query natural language text, e.g.: SearchValueSet, SearchCategory, SearchConcept

The functions described in table 3 are / will be used by ECDC client applications to populate certain fields in their interactive forms; as source for (semi)automated meta-data tagging; to enable 'time machine' functionality allowing to use old (obsolete) versions of terms while working with 'old' but still scientifically valuable content. (In epidemiology, data and information is re-used over decades, e.g. data on 'Spanish flu' is reused in studies of understanding current influenza pandemic.)

4 Discussion and Conclusions

Why to build and use an in-house terminology service, while there are a number of well managed, comprehensive sources out there? Although these sources are very valuable in providing references and standard approach in designing structure and functions, obviously they will never answer to the specific internal needs of a given organization. We need navigation within our own complex term space with multiple views, answering conflicting granularity needs, clustering along different perspectives specific to ECDC, etc.

The experiences to build up a shared terminology service for ECDC taught us, that such an approach (setting up a standard based, ontology 'enhanced' terminology server) has no serious technical, IT engineering obstacles these days. The existing standards on different levels allow a straightforward, transparent approach, so far the developers of the client systems were able to interpret and use (parse) the messages of the TS without significant problems. Lessons learned:

(1) Both to achieve internal consensus on and to build up shared terminology contents needed significantly more resources than planned at the beginning.

(2) Existing external reference term sets proved to be relatively poor regarding the domain of public health.

(3) Although using SKOS for value sets and OWL for the ontology caused some problems in how to build up the bindings and navigating functionality among the two levels, this has been solved by some restrictions on what OWL constructs could be used.

We think that the chosen approach, based on W3C rather than CEN TC 251 or ISO TC 215 standards, might be future proof and adds practical interoperability. We hope that by publishing the public health and communicable disease related terminology services we will help to fulfill the mandate of ECDC in assisting member states. This work triggered also several interoperability efforts on lower level of system to system messaging and also on higher level of services alignment. Further steps in utilizing the terminology services in building semantic search and knowledge navigation tools will follow.

References

1. Cimino, J.: Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inf. Med. 37(4-5), 394–403 (1988)
2. Goldfarb, C., Rubinsky, Y.: The SGML handbook. Oxford University Press, Oxford (1990)

3. Balkányi, L.: Terminology services: an example – an example of knowledge management in public health. Euro Surveill. 12(22) (2007)
4. Public Health Information Network Vocabulary Access and Distribution System, http://phinvads.cdc.gov/vads/SearchVocab.action
5. Family of International Classifications, http://www.who.int/classifications/en/
6. Medical Subject Headings, http://www.nlm.nih.gov/pubs/factsheets/mesh.html
7. Open Biomedical Ontologies, http://www.obofoundry.org/
8. CEN TC 251, http://www.cen.eu/CENORM/Sectors/TechnicalCommitteesWorkshops/CENTechnicalCommittees/CENTechnicalCommittees.asp?param=6232&title=CEN%2FTC+251, ISO TC 215, http://www.iso.org/iso/iso_technical_committee?commid=54960
9. Rodrigues, J.M., Kumar, A., Bousquet, C., Trombert, B.: Standards and biomedical terminologies: the CEN TC 251 and ISO TC 215 categorial structures. Stud. Health. Technol. Inform. 136(issue), 857–862 (2008)
10. IHTSD, http://www.ihtsdo.org/
11. Health Level Seven, http://www.hl7.org/
12. HL7, problem areas, http://hl7-watch.blogspot.com/2005/11/list-of-problem-areas.html
13. SPARQL Protocol and RDF Query Language, http://www.w3.org/TR/rdf-sparql-query/
14. Simple Knowledge Organization System, http://www.w3.org/TR/2009/CR-skos-reference-20090317/
15. Ontology Web Language, http://www.w3.org/TR/owl-features/
16. DOLCE: http://www.loa-cnr.it/DOLCE.html
17. Rosse, C., Mejino Jr., J.L.V.: A Reference Ontology for Biomedical Informatics: the Foundational Model of Anatomy. J. Biomed. Inform. 36(6), 478–500 (2003)