Process of Building a Vocabulary for the Infection Domain

Gayo Diallo[1], Patty Kostkova[1], Gawesh Jawaheer[1], Simon Jupp[2], Robert Stevens[2]
[1]*City eHealth Research Centre, City University, London, UK*
[2]*School of Computer Science, University of Manchester, Oxford Road, Manchester, UK. M13 9PL*

first.last.1@city.ac.uk, patty@soi.city.ac.uk, first.last@manchester.ac.uk

Abstract

The Semantic Web vision relies on metadata and semantic annotation to be implemented on real world data. Ontologies and ontology-like artefacts are the key component providing necessary knowledge for web document description. Domain ontology building is, however, a difficult and time consuming task. In this paper, we present our process of building an infection domain vocabulary for the National electronic Library of Infection. This paper describes the requirements for the vocabulary development process and the initial results.

1. Introduction

With the proliferation of information available on the Internet, it becomes increasingly difficult to provide relevant information to users. Generalist search engines do not cater for the diverse variety of users and user groups with different preferences, information needs and priorities. Key word searches with available search engines are often good enough to find results, but all too frequently do not account for the inherent ambiguity of terminologies used in diverse domains.

We have opted to take this reality into account for the National electronic Library of Infection (NeLI[1]), by exploiting Semantic Web technologies [1]. The Semantic Web activity aims to introduce meaning to the Web using semantics delivered by ontologies and other knowledge organisation systems (KOS). An ontology or KOS provides a shared understanding of a domain and can be used to semantically characterise online resources through annotation.

In this paper we discuss the requirements and design of an infection domain KOS for the NeLI portal referred sometimes as the NeLI ontology. NeLI is a real-world Internet portal providing the best evidence base for professionals in infection domain developed in collaboration with the key stakeholders and domain experts. This KOS is intended to be used for the annotation of NeLI resources and the personalisation of the NeLI portal by providing customised services to users. The NeLI portal is accessed by a number of medical professionals with different preferences and medical information needs.

The paper is organised as follows. The next section gives an overview of the NeLI background and a short overview on related work is presented in section 3. We describe the requirements for the NeLI KOS in section 4. In section 5 we present the resources used in the building process together with the method employed. Before concluding, we present our initial results in section 6.

2. NeLI background

The National electronic Library of Infection (NeLI) provides a single access point to the best available evidence on all aspects of infection. The family of NeLI projects include the National Resource for Infection Control (NRIC, www.nric.org.uk), Bugs and Drugs on the Web (www.antibioticresistance.org.uk) and other infection-related projects. Funded by the UK's Department of Health and the Health Protection Agency, NeLI provides the main source of online medical evidence for a wide spectrum of users - clinicians, family doctors (GPs), environmental health officers, infection control nurses, and other public health professionals.

The main sources of medical evidence are books, journals, and Internet-based sources. These include: The Health Protection Agency portal (http://www.hpa.org.uk), Cochrane database

[1] http://www.neli.org.uk

(www.cochrane.co.uk), NHS Centre for Reviews and Dissemination, Effective Health Care Bulletins, British National Formulary, CDC, BMJ, WHO, Department of Health and others. The quality and reliability of the information provided does, however, vary significantly. Although readers rely on journal review articles and editorials, the scientific evidence of these is inherently unreliable and biased towards a positive and optimistic view of the effectiveness of intervention. The key value-added feature distinguishing NeLI from other medical portals is the aim to provide, for each article, quality-appraised, evidence-based knowledge with clearly identified the level of evidence of its study.

3. Knowledge Organisation System (Vocabulary) building

Building an ontology or KOS is not easy. In contrast to the database domain, there is no commonly agreed methodology. The process, which usually involves domain experts, must lead to a clear, coherent, easy to use and extensible artefact [2]. Ontology, terminology and other KOS building projects have been undertaken for many decades [3-10]. Some methodologies start from scratch while others reuse existing vocabularies or ontologies. More recently, some (semi-)automated building methodology from texts emerged [11-14], as the majority of any organisation's knowledge is encoded within the textual documents. The general idea is to identify the most important concepts within the texts as well as hierarchical and transversal relationships using Natural Language Processing (NLP) and Information Extraction (IE) tools, sometimes using external knowledge resources.

Once designed, the ontology or KOS has to be implemented in a specific language. Editor tools constitute help for this task when all choices have already been made. Protégé [15] is widely used for ontology building. The tool is regularly evolving and thanks to the incorporation of new plugins, users are benefiting frequently from new functionalities. Protégé allows the description of an ontology in formats suitable for its use within the Semantic Web infrastructure. RDF/RDFS [16][17] and the Web Ontology Language (OWL) [18] are currently the most representative. The Simple Knowledge Organisation System (SKOS) [19] is, however, a recent semantic language development that is based on RDFS and it is destined to play a key role in semantic description. SKOS can be used for the representation of terminologies, vocabularies and other concepts schemes. Such KOS lack the formal, ontological distinctions made in many ontologies used in their pursuit of describing the nature of entities in the world and information about the world. KOS, on the other hand, often simply describe looser conceptual descriptions of the "relatedness" of terms within a domain or how a particular domain is organised by its members, whether or not such an organisation has much bearing on reality. Such KOS are, however, extremely valuable in information retrieval, indexing, information navigation, etc.

4. Requirements for the NeLI ontology

In this section we describe the requirements and usage for the NeLI KOS. The KOS or vocabulary will cover the wide spectrum of infection, including clinical microbiology, clinical infectious disease, infection control, public health, surveillance and populations affected. The infection domain includes the investigation, diagnosis, treatment and control of infectious diseases at a clinical and public health level. From a functional requirement point of view, there is no need to provide clinical diagnostic and decision making support as the library consists of the best available evidence interpreted by doctors rather than providing a clinical decision support system or linking to patient electronic records. Therefore, formal reasoning for this particular need is not our primary goal. Instead, the ontology is used for navigation, search and quick access to customised resources for groups of users. We describe below the usage scenarios of the NeLI KOS.

Semantic annotation of infection resources. In order to support a user-customisable search, resources need to be precisely described. Semantic annotation is the process of attaching one or more semantic entities to a resource. Resources managed through the NeLI portals will be both characterized by the NeLI KOS and Dublin Core [20] which defines a list of fields characterising an electronic document for cataloguing and search purposes. This semantic annotation process performed manually is a difficult, time consuming task specifically when one has to deal with a very large knowledge source (several thousands of entities). Infection control resources would be tagged on NeLI (semi)automatically by the relevant bacteria, disease and population. However, in NRIC (an infection control specific portal hosted by NeLI) the same document would be tagged by bacteria, disease, mode of transmission, healthcare setting where it is to be controlled etc. That is why the KOS has to include the maximum possible terms to cover the domain.

These requirements can be summarised as indexing, exploration, navigation and query for information retrieval. This suggests a KOS style vocabulary would be appropriate. Such ontology-like artefacts can be represented in SKOS, which offers access to many facilities of the Semantic Web, but with a lesser (though significant) effort than is required for a formal ontology. This choice also allows one of the largest such KOS, the Medical Subject Headings (MESH) in its SKOS representation to be available for use.

Personalisation of the NeLI portal. A "one fit for all" approach for searching and ranking discovered knowledge on the Internet does not cater for the diverse variety of users and user groups with different preferences, information needs and priorities. One of the key factors for accurate and effective information access is, however, the user context. Allan et al. [21] define contextual retrieval as the combination of search technologies and knowledge about query and user context into a single framework in order to provide the most appropriate answer for a user's information needs. Even if the above definition focuses on information retrieval, the user's information context is important for a wide range of domains including semantic browsing and filtering. Sieg *et al*. [22] consider that the critical elements that make up a user's information context include the semantic knowledge about the domain being investigated, the short-term information need as might be expressed in a query, and the user profiles that reveal long-term interests.

A recent survey on the NeLI portal web logs shows that users prefer the browsing modality to other information access modalities [23]. The current NeLI portal front-end uses a static tree identified by infectious disease experts as a controlled vocabulary for searching. The tree contains the entries judged most relevant and reflects the needs of the communicable disease domain. In the future, the tree will be dynamically built by pruning the NeLI KOS according to the context of the current user. Moreover, for users preferring to search the portal the KOS will help by suggesting contextual keywords.

Browsing of Infection domain related resources. The browsing issue is motivated by the development of the EU funded project SeaLife [24]. The main goal of the project is to develop a grid browser for the life sciences, which will link the current web to the current emerging e-science infrastructure. Particularly, SeaLife will help users to semantically browse the web by highlighting the semantic data contained in the web page which that is being visited and providing a dynamic access to web servers and Web/Grid services related to the web content retrieved. The system is built using the Conceptual Open Hypermedia Service (COHSE) system [25] developed at The University of Manchester. COHSE automatically adds hyperlinks on web pages by recognizing terms contained in background knowledge, based on an ontology or KOS. For any term highlighted on a page, resources are provided for broader, narrower and related terms obtained from the underlying vocabulary. The NeLI KOS will be used by a dedicated version of the COHSE system in order to provide access to resources and services specific to the infection domain. Users searching NeLI often need additional information, definitions of diseases, vaccination, drugs dosages, etc. that they have to search for manually. Thus, by adapting COHSE we will be able to provide additional information on viewing resources on NeLI using SeaLife/COHSE semantic mapping to appropriate online portals and databases (called targets). SeaLife's component COHSE uses the NeLI KOS to highlight the concept labels (both preferred and alternative) keywords for each user group on the web page that is being viewed. These highlighted terms are used as insertion points for dynamically allocated hyperlinks. On clicking a COHSE dynamic hyperlink, the user is offered a range of services (other pages, searches specific to that term, etc.) suitable for that page. The user can, however, use the broaderthan, narrowerThan and related to links of the KOS to move to more specific or general targets, depending on their task. Then, for example, searches can be carried out on targets to provide users with additional information about antibiotics for a particular infection based on the terms available for the concept for that infection.

From the scope of the NeLI KOS described in this paper, we have clearly identified a set of requirements that must be fulfilled. Firstly, the *need for explicit and natural language naming of concepts*. This makes the KOS understandable for the information officers of NeLI as well as other external human users. Further we need to provide a free text definition of the concept. Secondly, we need *multiple labels for concepts* in order to cope with synonyms, preferred and the alternatives terms such as allowance for English and American spellings, abbreviations, etc. Finally we need to *map the ontology to other biomedical vocabularies/ontologies to allow* interoperability and to gather additional information and services while browsing.

5. Existing resources and practical implementation

5.1 Resources and tools

5.1.1 The MeSH vocabulary

Initial work was conducted by medical domain experts working on NeLI, in close collaboration and consultation with the NeLI Chair and Advisory Board members with representations from all major UK societies in infection and public health. The MESH vocabulary was identified as the most widely used indexing tool for medical libraries.

Medical Subject Headings (MeSH) is a controlled vocabulary maintained by the U.S. National Library of Medicine[2]. It is mainly used for annotating and indexing articles from PubMed.

The MeSH vocabulary provides a consistent way to retrieve information that may use different terminology in different articles for the same concepts. MeSH is organized in a directed graph, with concepts such as anatomy and diseases, but also geographic locations, at the top level.

The MeSH vocabulary is used for indexing journal articles from Index Medicus and Medline. It also provides access and links to the integrated molecular biology databases maintained by the National Centre for Biotechnology Information.

5.1.2 Domain experts inputs

The MeSH vocabulary has been judged inappropriate for indexing NeLI. Whilst the structure of MeSH is appropriate and its linguistic representations are very useful, some areas of NeLI's domain are not covered at all. Therefore, in a first stage terms relevant for infection and public health were extracted by pruning MeSH, and then additional sections on populations, health organisations and public health provided by domain experts were added to this MeSH fragment meet the requirements of the scope of the domain. These were sent for consultation to the Advisory Board members and the final version, with taxonomical relationships served as the indexing and navigation graph for the first version of the library.

In parallel, NeLI launched NRIC, a portal dedicated to infection control that needed a taxonomy specific to infection control – this was developed by the NRIC content manager and domain expert in a similar way as the broader NELI KOS was done previously but including many specific terms, such as the modes of transmission and disinfection techniques.

At the second stage, experts and users started suggesting new terms and contributed to the evaluation of new versions of the NeLI/NRIC KOS as the portals were expanding. Both systems were merged into a single tree and migrated into OBO format – still as a taxonomical tree without semantic non-taxonomical relationships and with ever growing need for evolution.

5.1.3 The SKOS language and plugin editor for Protégé 4

Protégé 4[3] is an editor for OWL ontologies. It is being developed as a joint collaboration between The University of Manchester and Stanford University. The SKOS plugin[4] for Protégé 4 offers support for viewing and editing SKOS vocabularies. SKOS represents a family of formal languages providing a model to represent and use vocabularies and ontologies in the framework of the Semantic Web. A SKOS vocabulary is composed of Concepts that belong to Concept Schemes. A Concept can be simply defined as a "Unit of thought". Concepts have properties, there are two major types of properties used in SKOS. The first type are semantic relations and are used to relate concepts to each other giving structure to the KOS, one such property is `skos:broader` which can be used to assert that one concept is broader in meaning, or less specific, to another concept. The second type provide labels, descriptions, comments etc to a particular concept, the `skos:prefLabel` and `skos:altLabel` properties can be used to specify preferred and alternative lexical labels for concepts.

5.2 Practical Implementation

The first step was to convert the initial NeLI ontology, represented in OBO format, to SKOS. The OBO version of NeLI is converted into SKOS and represented using RDF/XML syntax. Each term from OBO becomes a SKOS concept and a member of that concept scheme, full details of the mapping from OBO to SKOS can be found here[5].

NeLI, NRIC content managers and information specialists and collaborating domain experts were given the SKOS editor and jointly worked on adding definitions and adjusting relationships to come to an agreed structure for the KOS over the phone, VNC and

[2] www.nlm.nih.gov/

[3] http://protege.stanford.edu/

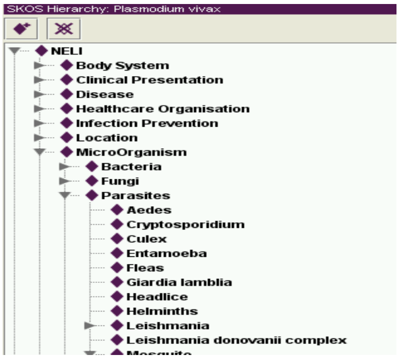
[4] http://code.google.com/p/skoseditor/

[5] http://www.cs.man.ac.uk/~sjupp/skos/index.html

at a face-to-face workshop. This version of the NeLI ontology includes the core infection domain and public health including entities relationships such as *is caused by* (and its inverse *causes*) to relate an infection disease to its micro organism causing agent, *is treated by* (and its inverse *treats)* to relate a disease to its treatment.

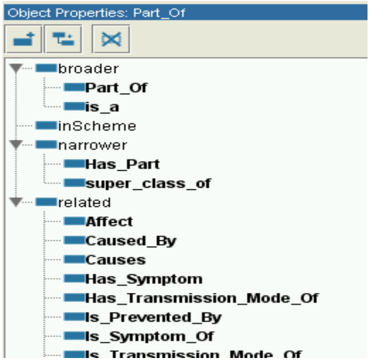
6. Initial results

The KOS has 630 concepts organised around 12 Top entities. They include Population, Prevention of Infection, Symptom, Transmission Mode and Treatment. Figure 1 gives an overview of the hierarchy edited under the SKOS editor plugin. Each concept of the hierarchy has an ID, a preferred label, one or more alternative labels and a definition in natural language text. It may also have one or more relationships to other concepts of the hierarchy.


SKOS Hierarchy: Plasmodium vivax

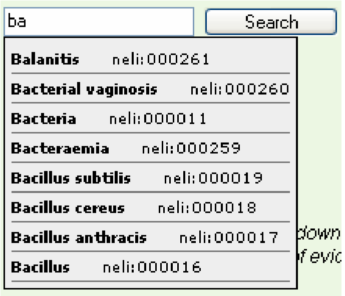
- NELI
  - Body System
  - Clinical Presentation
  - Disease
  - Healthcare Organisation
  - Infection Prevention
  - Location
  - MicroOrganism
    - Bacteria
    - Fungi
    - Parasites
      - Aedes
      - Cryptosporidium
      - Culex
      - Entamoeba
      - Fleas
      - Giardia lamblia
      - Headlice
      - Helminths
      - Leishmania
      - Leishmania donovanii complex
      - Mosquito

Figure 1. Overview of the NeLI ontology edited under the SKOS editor plugin.


Object Properties: Part_Of

- broader
  - Part_Of
  - is_a
- inScheme
- narrower
  - Has_Part
  - super_class_of
- related
  - Affect
  - Caused_By
  - Causes
  - Has_Symptom
  - Has_Transmission_Mode_Of
  - Is_Prevented_By
  - Is_Symptom_Of
  - Is_Transmission_Mode_Of

Figure 2. Overview of the semantic relationships of the NeLI ontology


ba

Search

Balanitis	neli:000261
Bacterial vaginosis	neli:000260
Bacteria	neli:000011
Bacteraemia	neli:000259
Bacillus subtilis	neli:000019
Bacillus cereus	neli:000018
Bacillus anthracis	neli:000017
Bacillus	neli:000016

Figure 3. Recommended keywords retrieved from the NeLI ontology

The concepts of the hierarchies are related by a set of semantic relationships (Fig. 2).

The first usage of the NeLI KOS has been to provide an auto-completion service that suggests to users keywords to be used during the search process.

When typing a search term, a query is sent to the NeLI server using Asynchronous JavaScript and XML (AJAX). A list of terms is sent back to the user and displayed in a combo box (Figure 3). Instead of providing the same list to different users (e.g., nurses or clinicians) having different needs, user profile can be taken into account when querying the server and returning the result.

7. Conclusion

We have described in this paper our ongoing work on the designing and building of a domain vocabulary or KOS for infection domain related resources annotation and browsing as well as personalising the portal of the National electronic Library of Infection. We have described the requirements for the semantic support and argued that a KOS, rather than a formal ontology, suits those requirements and we have presented the first results of that choice. One of the uses of the ontology is to automate the annotation of resources. However, we noted that it is not straightforward to capture the way the information officers organise resources. For example, in the NRIC portal resources are annotated by modes of transmission (blood-borne, food-borne etc) which is not found in the medical resources but uses the tacit knowledge of the domain expert. We argue also that the process of ontology building can benefit from the advantage offered by the Semantic Web technologies as people can collaboratively create and build common vocabulary without centralized control. For example, we have used a SKOS representation for MeSH and extended that with concepts and terms appropriate for NeLI, while both keep their separate identity. Further investigation is needed to build suitable infrastructures supporting the possibility for people geographically disseminated, to argue for example on every assertion made during the building process. Future work includes also improving the KOS by the NeLI portal web logs processing.

8. Acknowledgments

We acknowledge the European Commission for providing funding for the SeaLife project (IST-2006-027269). In addition, we would like to thank Sue Wiseman, Faiza Hansraj and Jane Mani-Saada for the domain expertise in developing the ontology and the NeLI advisor board members chaired by Roger Finch for their input and ongoing help in the development process. Also, we would like to acknowledge the HPA SLI team and the HPA library service, in particular, Margaret Clennett and David Keech, for their expertise and ongoing input into the ontology development process.

9. References

[1] Berners-Lee, T., Hendler, J., Lassila, O. "The Semantic Web", *Scientific American* 284(5):34-43 (May 2001).

[2] Guarino, N. (1998). Formal ontologies and information systems. *In FOIS'98*, Trento, Italy, IO Press.

[3] Gruber, T. Towards principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies*, 43 (5/6):907-028.

[4] Lenat, D.B.,Guha, R. V., Building large knowledge based systems. Reading, Massachusetts: Addison Wesley. (1990).

[5] Grüninger, M., Fox, M. S., Methodology for the design and evaluation of ontologies. *IJCAI Workshop on Basic Ontological Issues in Knowledge Sharing*, Montreal, Canada. (1995).

[6] Fernandez, M., Gomez-Perez, A. et Juristo, N. (1997). Methontology : From ontological art towards ontological engineering. *Spring Symposium Series*, pages 33_40.

[7] Uschold, M., King, T. Towards a methodology for building ontologies. In Proceedings IJCAI-95, Workshop on Basic Ontological Issues in Knowledge Sharing, Canada.

[8] Gomez-Perez, A. et Rojas, M. Ontological reengineering and reuse. In European Knowledge Acquisition Workshop (EKAW). (1999).

[9] Benjamins, R. et Fensel, D. (1998). The ontological engineering initiative-ka. *Proceedings of the 1st International Conference on Formal Ontologies in Information Systems, FOIS'98*, Trento, Italy, pages 287-301.

[10] Jarrar, M., Meersman, R. Formal ontology engineering in the dogma approach. *Proceedings of the International Conference on Ontologies, Databases and Applications of Semantics (ODBase 02)*, LNCS 2519:1238 _ 1254. (2002).

[11] Biebow, B., Szulman, S., Terminae : A linguistic-based tool for the building of a domain ontology. *11th European Workshop, Knowledge Acquisition, Modeling and Management (EKAW' 99)*, Dagstuhl Castle, Germany, pages 49-66.

[12] Maedche, A., Staab, S., Semi-automatic engineering of ontologies from text. *Proceedings of the Twelfth International Conference on Software Engineering and Knowledge Engineering (SEKE'2000).* 2000.

[13] Fortuna, B., Grobelnik, M., Mladenic, D., Semi-automatic Data-driven Ontology Construction System. *Proceedings of the 9th International multi-conference Information Society IS-2006*, Ljubljana, Slovenia.

[14] Sanchez and Moreno, Sanchez, D., Moreno A., A methodology for knowledge acquisition from the web. In *International Journal of Knowledge-Based and Intelligent Engineering Systems*, 9 1-23. IOS Press. 2006.

[15] Gennari, J., Musen, M. A., Fergerson, R. W., Grosso, W. E., Crubezy, M., Eriksson, H., Noy, N. F., Tu, S. W., The evolution of protege: An environment for knowledge- based systems development. *International Journal of Human-Computer Studies*, 58 :1:89-123. (2003).

[16] Lassila, O., Swick, R. R., Resource description framework (rdf) model and syntax specification. *World Wide Web Consortium W3C Recommendation* 22 February 1999 /

[17] Decker, S., Mitra, P. et Melnik, S., Framework for the semantic web: An rdf tutorial. 4(6):68-73, 2000

[18] McGuinness, D. L. et van Harmelen, F. (2004). Owl web ontology language overview. World Wide Web Consortium, Recommendation REC-owlfeatures-20040210.

[19] Miles, A., Brickley, D., Skos core guide. Technical Report, *2nd W3C Public Working Draft* 2 November 2005.

[20] Hillman, D. (2001). Using Dublin Core. DCMI Recommandation. http://dublincore.org/documents/usageguide/

[21] Allan, J. et al. Allan, J. et al.(2003) Challenges in Information Retrieval and Language Modelling : Report of a workshop held at the centre for intelligent information retrieval, University of Massachusetts Amherst, September 2002. ACM SIGIR Forum, 37(1):31-47.

[22] Sieg, A., Mobasher, B., Burke, R., Representing Context in Web Search with Ontological Users Profiles. *Proceedings of the Sixth International and Interdisciplinary Conference on Modeling and Using Context (Context'07)* Lecture Notes in Artificial Intelligence, Vol. 4635, PP. 439- 452, Springer-Verlag, Berlin, 2007.

[23] Roy, A., Kostkova, P., Carson, E., Catchpole, M., Web-based provision of information on infectious diseases: a systems study, Health Informatics J. 2006 Dec;12(4):274-92.

[24] Schroeder, M., Burger, A., Kostkova, P., Stevens, R., Habermann, B., Dieng-Kuntz. R., From a Service-based eScience Infrastructure to a Semantic Web for the Life Sciences: The Sealife Project. In *Workshop on Network Tools and Applications in Biology, NETTAB 2006*, Sardinia, Italy, 2006.

[25] Bechhofer, S., Yesilada Y., Horan, B., *COHSE: Knowledge-Driven Hyperlinks* the Semantic Web Challenge at the International Semantic Web Conference (ISWC 2006), 2006.