KR-MED 2006 "Biomedical Ontology in Action"
November 8, 2006, Baltimore, Maryland, USA

An Online Ontology: WiktionaryZ

Erik M. van Mulligen, Ph.D.[1,2], Erik Möller, Peter-Jan Roes[3], Marc Weeber, Ph.D.[2], Gerard Meijssen, Christine Chichester Ph.D.[2,4] , Barend Mons Ph.D.,[1,2,4]
[1]Dept. of Medical Informatics, Erasmus Medical Center, Rotterdam, the Netherlands
[2]Knewco Inc, Rockville, United States of America
[3]Charta Software, Rotterdam, the Netherlands
[4]Human and Clinical Genetics, Leiden University Medical Center, Leiden, the Netherlands
e.vanmulligen@erasmusmc.nl

There is a great demand for online maintenance and refinement of knowledge on biomedical entities[1]. Collaborative maintenance of large biomedical ontologies combines the intellectual capacity of millions of minds for updating and correcting the annotations of biomedical concepts with their semantic relationships according to latest scientific insights. These relationships extend the current specialization and participation relationships as currently exploited in most ontology projects. The ontology layer has been developed on top of the Wikidata[2] component and allows for presentation of these biomedical concepts in a similar way as Wikipedia pages. Each page contains all information on a biomedical concept with semantic relationships to other related concepts. A first version has been populated with data from the Unified Medical Language System (UMLS), SwissProt, GeneOntology, and Gemet. The various fields are online editable in a Wiki style and are maintained via a powerful versioning regiment. Next steps will include the definition of a set of formal rules for the ontology to enforce (onto)logical rigor.

INTRODUCTION

In order to deal with the deluge of biomedical information many projects have been initiated that aim at semantically annotating content. Many of these projects can be characterized as an attempt to exploit advanced natural language processing and text mining technology to identify the relevant semantic topics contained in a text[3]. By identifying these concepts in a text one can exploit available information about a concept as being formalized in an ontology for a number of tasks. One of these tasks is to improve information retrieval[4] (e.g., retrieval of texts on a particular concept might also include the retrieval of documents with a more specific, narrower meaning). Another task would be semantic navigation between texts (e.g., exploring the semantic relationships between an identified concept in a text and concepts in other texts[5]).

Outside the biomedical domain the W3C has been working on defining exchange standards for ontologies. Their objective is to facilitate the development of technologies that enable cross-community data integration and collaborative efforts by adding semantics to the data. An example is the semantic web where webpages are semantically tagged and through these semantic tags linked to other webpages (similar to the current hyperlinked web). RDF, OWL and DAML[6] are examples of standards to impose semantic tags on information on the web. The meaning of these tags is captured in ontologies that contain additional information on how these semantic tags interrelate. These semantic interrelated tags can be used by applications for instance to semantically navigate between web resources.

All these tasks heavily rely on ontologies that serve as a repository of these biomedical concepts. Ontologies provide facilities to semantically relate the different biomedical topics. A first generation of ontologies (with limited scope) is available now. Good ontological principles have been a research topic and many scientific projects aim at a next generation of ontologies[7]. The Open Biomedical Ontologies consortium provides a platform for making available ontologies for shared use in the medical and biomedical domain that have been constructed with tools that bring in a greater degree of logical and ontological rigor[8]. Various tools have

been constructed that assist users with constructing these ontologies. Protégé is a freely downloadable program to construct ontologies using a strong formalism[9].

OntoBuilder is another ontology editor that has been developed to automatically derive ontologies from a corpus (web pages) with support to refine and restructure them. Its focus is in particular on ontologies supporting the semantic web[10]. The main emphasis of all these tools is to make the development of (rigorous) ontologies easier. The whole process of collaboration, discussion and interrelating ontologies has not yet been addressed in these tools.

In this paper a mechanism is presented to harvest from existing ontologies originating from different sources and make these ontologies available for web-based refinement through a collaborative effort of the community of scientists. The hypothesis is that the online interaction, discussion and annotation of biomedical concepts will lead to wider coverage and higher quality ontologies with more semantics defined. Typically, most ontologies limit themselves to defining a hierarchy containing the specialization or participation relations. The biomedical semantic relations (a particular biomedical concept has a particular semantic relationship with another biomedical concept) require experts to interact and refine. These are important for the next generation of intelligent applications.

It is clear that an ontology has to cover a substantial part of the domain in order to be useful. In the biomedical domain, this would require that at least a substantial part of all medical concepts and of all genomic and proteomic concepts have to be in. Current vocabularies in these fields yield about 1,352K concepts for the medical domain (UMLS[11]) and about 200K for the genomics and proteomics domain (Swiss-Prot, EntrezGene, and Gene Ontology[12]).

Building a comprehensive ontology is an enormous endeavor. Bringing together all ontological knowledge from different biomedical disciplines in one environment seems to be quite impossible. Furthermore, a biomedical ontology is not a static, one-time effort. Such an ontology should be continuously revised and updated with the latest new biomedical concepts and the latest semantic relations between the concepts[1]. Only imagining the rate with which genomics and proteomics data are produced yielding new information on genes and proteins it becomes clear that a comprehensive and up-to-date ontology is beyond the capabilities of any single scientific project.

The only way to cope with such enormous amounts of data in so many different biomedical fields is to have an open environment in which all scientists can collaboratively share their knowledge on particular biomedical topics. Therefore we are currently investigating the possibilities of using a web-based approach to build and maintain biomedical ontologies. Benefiting from the pioneering work of the Wikimedia Foundation on collaborative development of web-based encyclopedias, we are exploring the possibilities to adapt a Wikimedia product in such a way that it can be used to support collaboration on ontology work: the WiktionaryZ software.

Many of the current vocabularies do not satisfy the ontological principles as current research has defined[13]. In addition, editing and updating ontologies should follow rules that guarantee soundness and correctness of the ontology. Description logic in combination with the specification of a separate hierarchy along the specialization and participation relation could make it possible to automatically detect errors in the concept classification. The WiktionaryZ has been developed in such a way that such an additional hierarchy can be expressed.

In addition to creating a collaborative instrument for biomedical scientists, this approach is also of interest to language engineering scientists. A systematic translation of biomedical terms is a rich source for language engineers and of great interest to them.

METHODS

The architecture of WiktionaryZ (see Figure 1) has been based on the existing MediaWiki software. Wikidata itself is an extension of the MediaWiki

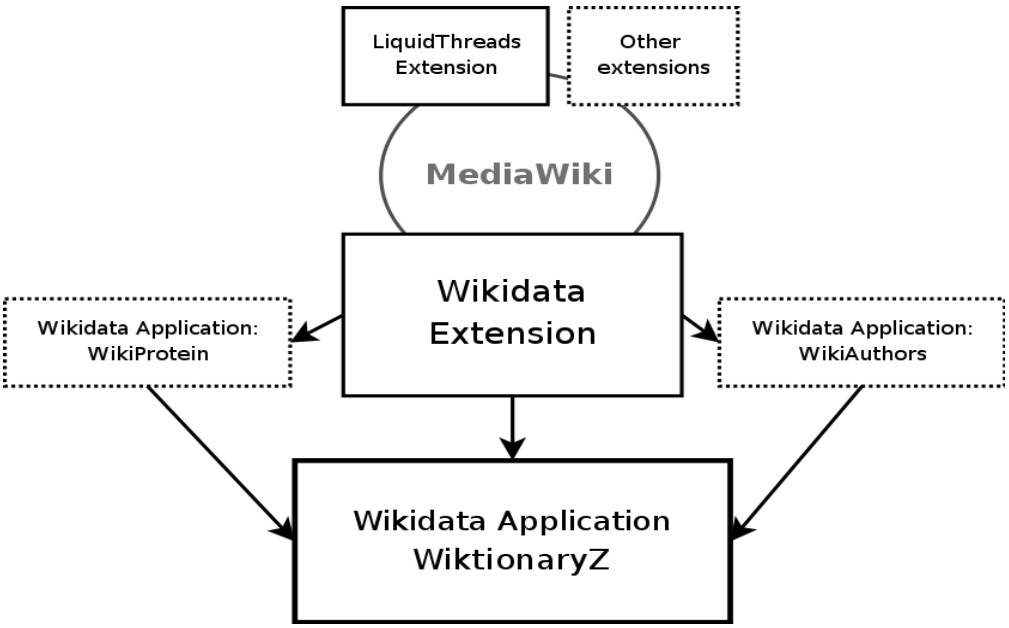


Figure 1 - Schematic overview of the architecture of WiktionaryZ. It has been developed on top of the existing MediaWiki software.

software that allows for structured data functionality beyond editing flat documents like Wikipedia articles. All data are stored in a MySQL relational database management system. WiktionaryZ has been built using Wikidata to store multilingual ontologies. It supports the notion of concepts, terms, synonyms, translations, definitions and alternative definitions, semantic relations, attributes, ontology class membership, and source annotations. Each of these elements is stored in the database as a separate entity. These entities can be combined in various queries supporting different applications. Specific applications (e.g., WikiProtein and WikiAuthors) can be defined as an implementation of the WiktionaryZ schema definition (with possibly some application-specific extensions).

The WiktionaryZ software provides the same functionality as the MediaWiki software with respect to online editing (talk pages) and version management. In order to distinguish between the ontology as provided by the authority - i.e. the organization that developed the thesaurus or vocabulary - and the version as maintained by the community an extended version management system is in place. The WiktionaryZ software discriminates between two version branches: the so-called authoritative version and the community version. These two branches are more or less independent: new versions of the authoritative version can be imported without disrupting the community version. Vice versa are edits made by the community clearly (visually) distinguishable from the authoritative version avoiding any confusion with respect to accountability. The authority can monitor and selectively include community edits to refine its own authoritative version. The community can harvest from the latest release of the version maintained by the authority after its import into the authoritative branch.

Every scientist can contribute and discuss information on a concept. The version management layer treats every edit as a new version. Versions can be rolled back if such a rollback does not cause relational inconsistencies. The LiquidThreads extension supports multiple threads per Wiki page. This means that one could have a discussion thread around the definition of a concept and a separate one for the translations of terms. The WiktionaryZ software and its database are available under a free content license as defined by the Free Content Definition (http://www.freecontentdefinition.org).

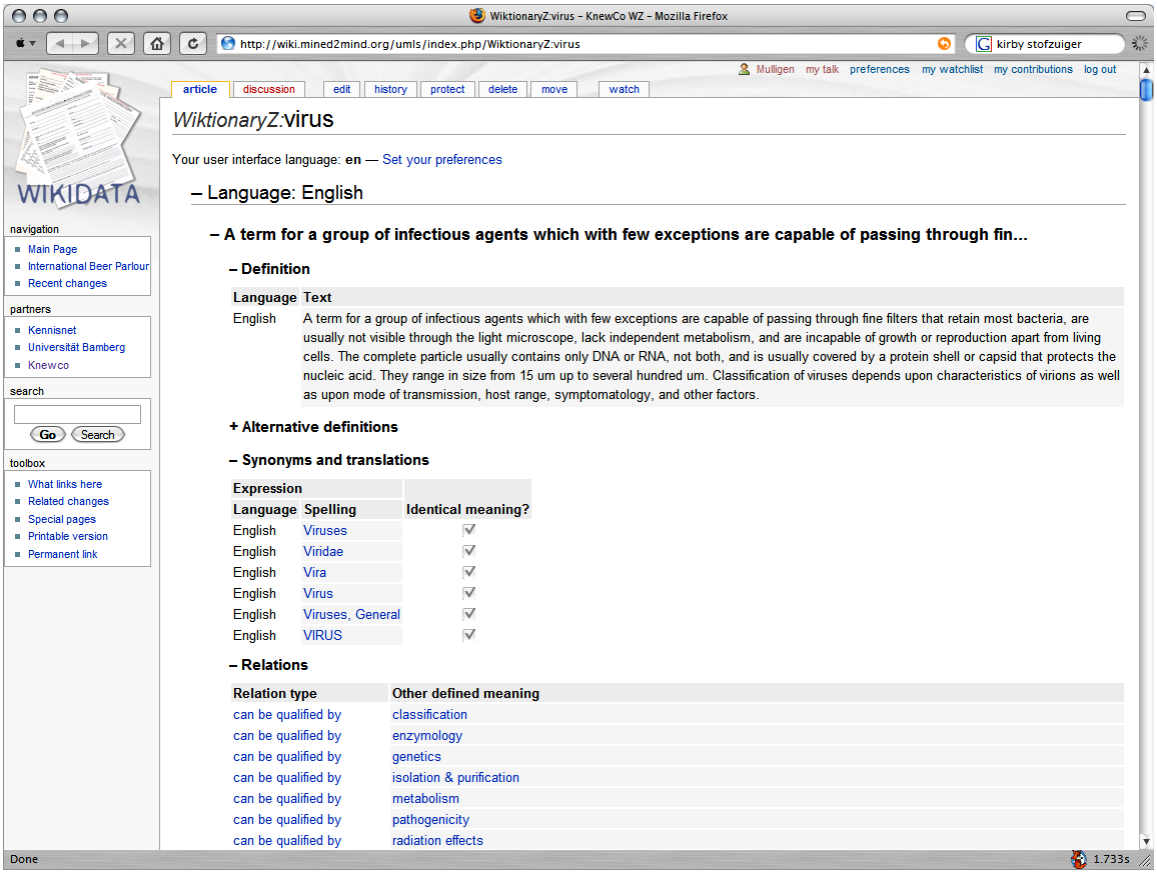
A Wikidata application is defined by a namespace and associated functionality. Each different vocabulary can have its own namespace and attached

to its namespace can be additional tables that require specific functionality. For instance, in the WikiProtein namespace each protein can be described by its own specific features, such as amino acid sequence, the species of origin, the experimentally identified function, etc. For a gene concept, the DNA sequence could be given. Despite these specializations for each namespace, the concepts share a common set of data (and structure) for each concept.

Each biomedical concept is defined by a definition – a short and precise specification of the concept. A biomedical concept can have additional definitions: these definitions might comprise real alternatives for the definition or definitions with a slightly different perspective: aiming at a different scientific discipline or at a different community (high school students, for instance). Figure 2 shows an example of the information comprised at a WiktionaryZ page. The palette of semantic relations between the biomedical concepts has initially been defined as the set of relations defined in the Semantic Network of the Unified Medical Language System[11]. This set of hierarchically organized relations can be easily extended and refined by the user.

Attached to each concept are terms (and synonyms), the language utterances used to refer to the concept. These terms are organized per language. Translations for each term can be entered and the system has been predefined with codes as defined in the ISO/FDIS 639-3 standard. Attached to each definition can be attributes. Initially these attributes will specify properties on the defined meaning: for instance the semantic type (e.g., a disease, a gene, a finding, a chemical, etc.) of the biomedical concept.

In order to benefit from the biomedical concepts as already defined in existing vocabularies and thesauri batch import facilities have been developed for the WiktionaryZ. Import facilities are now available for the UMLS files, Swiss-Prot files, Gene Ontology files, and the Gemet files. Most information contained in these vocabularies and thesauri has been succesfully imported and made available in a WiktionaryZ environment.



DISCUSSION

No other online editing environment has been developed that supports collaboration of scientists on annotation and semantic refinement of an ontology. The currently available tools allow for development of ontologies along some ontology design principles. However, many scientists need to be involved to refine the ontologies to a fine granular conceptual level, to annotate the concepts, and to express the semantic relationships between concepts, in short, to represent and codify the continuous advances of scientific knowledge about any biomedical subject. For effective use of ontologies in biomedical applications it is crucial to go beyond the current foundational relations of ontologies and beyond the well established and consistently described concepts.

Our first experiments with building the WiktionaryZ demonstrate that it is quite feasible to have large sets of concepts contained in a Wikidata database. The web based interface is fast enough to retrieve the concepts and combine all concept related data dispersed in different tables to the user. Pages are referenced per term. In case of a homonymous terms the page shows all the concepts for which the term is defined. The concept page can be very long. Currently WiktionaryZ does not provide any mechanism to define views on the data. A simple first approach would be to only show data for the language(s) that the user has indicated. More advanced views that are depending on the nature of the user's task can also be foreseen (i.e., differentiate between annotators, scientists, students, ontology developers, translators, high school students, etc.).

The WiktionaryZ does provide a powerful search facility: it searches for exact matches and allows for partial matches, both in the expressions associated with each concept and in their definitions. Misspellings and phonetic search are not implemented yet. It is evident that the current implementation lacks the ontological framework that allows for more sophisticated and rigorous quality control. This is essential when various users with different skill levels in ontology development are editing the ontology. Inclusion of a set of proper and well-defined relations expressed in a formal way should yield a more robust and more consistent editing of the ontology. Violation of these editing rules should lead to alerts to the user but should not be prohibited. It is at the moment unclear how much of the potential inconsistency problems can be avoided by this framework.

The alignment of different vocabularies also requires special attention. How can identical concepts defined in different vocabularies be aligned (mapped to the same concept)? It is yet unclear how we can support automatic detection of (almost) synonymous concepts (e.g., "water" and "H_2O" as being equivalent but defined in different vocabularies). This aspect has been a topic of study for already quite some years and we will explore the possibilities that have been identified.

A comprehensive biomedical ontology that can be effectively used for a number of tasks (bioinformatics, clinical medicine) will contain at least 2 million biomedical concepts. This is a rough estimate based on combining the current available thesauri, taken into account the overlap and the amount of non-medical concepts together with those parts that are still missing. Currently the National Library of Medicine, the Swiss Institute for BioInformatics, and the Gene Ontology Consortium have, apart from providing their sources, expressed their interest in this effort. An online maintained ontology will provide mechanisms to improve their authoritative sources as well.

In order to be able to include other ontologies/ thesauri as well the development of a method that can both read and write ontologies expressed in a standard syntax (OBO, OWL) has to be developed. This would make it possible to easily include a wide range of ontologies that are currently available in this format. Furthermore, the export allows the source authorities to download the latest edits for inclusion in their local version of the source. The current implementation of the system shows that it is technically feasible to have all these thesauri combined in one WiktionaryZ environment. What the impact - both with respect to quality and performance - of a large scientific community will be on such an online ontology remains a topic of research and will be part of future evaluation studies.

References

1. Wang K. Gene-function Wiki Would Let Biologists Pool Worldwide Resources. Nature 2006; 439-534
2. Möller E. Wikidata: Wiki-Style Databases. Available from: http://mail.wikipedia.org/pipermail/wikitech-l/2004-September/025377.html
3. Nagao K., Shirai Y, Squire K. Semantic Annotation And Transcoding: Making Web Content More Accessible. IEEE Multimedia, 2001;8(2):69-81
4. Müller H-M, Kenny EE, Sternberg PW. Textpresso: An Ontology-Based Information Retrieval and Extraction System for Biological Literature. PloS Biology, 2004;2(11).
5. Buitelaar P, Eigner Th, Racioppa S. Semantic Navigation With VieWs. Proceedings of the Workshop on User Aspects of the Semantic Web at the European Semantic Web Conference. 2005.
6. Miller E. Weaving Meaning : An Overview Of The Semantic Web. Presented at the University of Michigan, Ann Arbor, Michigan USA, 2004
7. Smith B, Rosse C: The Role Of Foundational Relations In The Alignment Of Biomedical Ontologies. Proc. Medinf 2004. Amsterdam: IOS Press, 2004;444-8.
8. Available from: http://obo.sourceforge.net/main.html
9. Knublauch H, Fergerson RW, Noy NF, Musen MA. The Protégé OWL Plugin: An Open Development Environment For Semantic Web Applications. Third International Semantic Web Conference, Hiroshima, Japan, 2004.
10. Roitman H, Gal A. OntoBuilder: Fully Automatic Extraction And Consolidation Of Ontologies From Web Sources Using Sequence Semantics. Proceedings of the International Conference on Semantics of a Networked World (ICSNW), 2006
11. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med. 1993;32(4):281-91.
12. Bada M, Stevens R, Goble C, Gil Y, Ashburner M, Blake JA, et al: A Short Study On The Success Of The GeneOntology. J Web Semantics 2004;1:235-40.
13. Smith B, Ceusters W, Klagges B, Köhler J, Kumar A, Lomax J, et al. Relations In Biomedical Ontologies. Genome Biology 2005; 6(5)