Developing an Ontology for Research and Care of Intracranial Aneurysms

Holger STENZHORN [1], Martin BOEKER,
Stefan SCHULZ and Susanne HANSER
Institute for Medical Biometry and Medical Informatics,
Freiburg University Medical Center, Freiburg, Germany

Abstract. We describe our approach to developing a domain ontology in the scope of the European project @neurIST. In this project a complex IT infrastructure is being created with the aim of providing a better understanding and management of intracranial aneurysms. A major goal of the project is to link and integrate highly heterogeneous, distributed data and knowledge from various scientific disciplines to improve information access for personalized disease management and research discovery. This task is to be based on an ontology representing all necessary entities from the different disciplines within the project, like biomedicine or epidemiology, and which are on a large spatial and temporal scale. The concrete applications of the ontology comprise terminological control, text mining, annotation, and data mediation. The ontology design and development is based on the DOLCE upper ontology and other existing domain ontologies which are either linked or partly included, such as the Foundational Model of Anatomy for anatomical entities and the UMLS for definitions and lexical information.

Keywords. ontology engineering, intracranial aneurysm, terminology

Introduction

The European project @neurIST aims at developing an integrated IT infrastructure to provide a better understanding and management of intracranial aneurysms. One of its main objectives is to integrate data from disparate sources and various scientific disciplines (e.g., biomedicine and epidemiology) which are characterized by high fragmentation and heterogeneity both in terms of format as well as scale. The project brings together partners from a multitude of different domains including computer scientists, biomedical informaticians, clinicians and experts in biomolecular research. The envisaged benefits for the end users (i.e., clinicians, scientists, and patients) include ameliorated diagnosis support, treatment planning and an easier and quicker access to domain knowledge, provided by an integrated system.

One main activity within the project is the development of a description logic-based ontology for representing the relevant concepts associated with intracranial aneurysms (http://ontology.aneurist.org) [1]. This ontology has to integrate both medical entities like anatomical, pathological or medical procedures and biomolecular, epidemiological and hemodynamic entities. It is also committed to combine the various disease

[1] Corresponding Author: Holger Stenzhorn, Institute for Medical Biometry and Medical Informatics, Freiburg University Medical Center, Stefan-Meier-Straße 26, 79104 Freiburg, Germany; E-mail: holger.stenzhorn@uniklinik.freiburg.de.

descriptions levels (e.g., clinician, genetic, and epidemiologic views) with various information sources (e.g., literature, clinical databases, and terminologies). Some example usage scenarios are [2]:

- Clinical data encompassing electronic patient records and data gathered via a dedicated clinical reference information model
- Epidemiological data originating from regional data collections including databases of general practitioners as well as of clinical studies
- Experimental data from molecular biology experiments done with patient samples including microarray data and genetic association studies

It has been quite challenging to properly identify the concrete needs towards the ontology, i.e., what domain experts and system developers expect it to contain to be adequate for their contexts. To alleviate this issue, and in parallel to the actual ontology development, some simple tools and a web-based ontology browser have been created that give straightforward and adaptable views on the ontology.

1. Concept Identification

For our purposes, we see an ontology as an artifact describing a specific, well-delineated domain and introducing the proper definitions and descriptions for all relevant concepts and relations in that domain. The concepts are arranged in a taxonomy and associated via relations (properties) by applying formal axiomatic definitions to constrain their interpretation and well-formed usage. Additionally, textual descriptions explain the meanings and appropriate uses of concepts and relations in plain language for domain experts and system developers.

As first step we created a list of concepts we deemed to be relevant for describing the disease of intracranial aneurysms, the affected patients and the technical devices and data structures used in the process of description, therapy and prognosis.

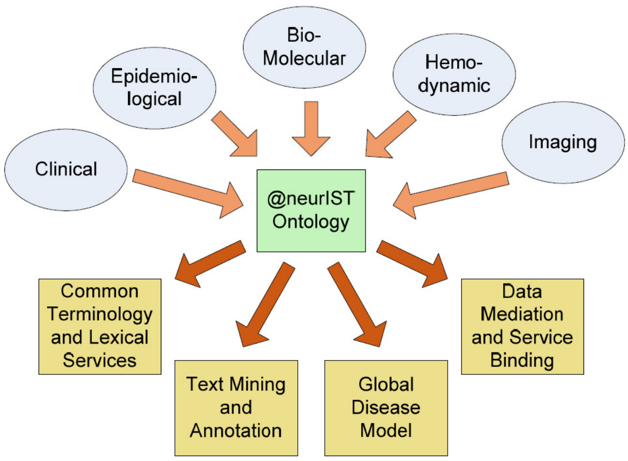

Figure 1. The various information sources integrated in the ontology and the tasks it is used for

As sources for the concepts, we surveyed all relevant data terminology and knowledge sources used in the project consortium (e.g., specific data base schemata, classifications used, disease models, information models, literature) (cf. Fig. 1). From theses concepts. an initial ontology skeleton was created covering and structuring the relevant conceptual space. At the same time, we interviewed and discussed with domain experts whether our modeling is correct and appropriate. But the above mentioned diversity within the project also led to diverse opinions among the experts on how to model certain concepts according to their respective contexts. To solve this issue, we divided the ontology in two basic parts, a "core" one that holds concepts classified in a context-independent fashion and a second context-specific, epistemological part. As an example, the smoking of tobacco is classified as a process in the context-independent part and in the second part it is classified within the clinical context as a known risk factor for developing an aneurysm rupture.

2. Ontology Design

2.1. Top-Level Ontology

We chose the Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) [3] as the basic, upper ontological framework and included its OWL DL implementation via the OWL import mechanism. We see DOLCE as ideal for our endeavor because of its "clear cognitive bias" and non-commitment "to a strictly referentialist metaphysics related to the intrinsic nature of the world". The root concept of DOLCE "particular" reflects a consensus that ontologies are systems of semantic types defining classes of particulars (individuals), and not concepts in the sense of entities of the mind. "Particulars" are further subdivided in terms of:

- Endurant – independent object- and substance-like wholes/entities, such as "Internal Carotid Artery" or "Blood"
- Perdurant – event-, process-, activity-, and state-like entities, like "Angioplasty" or "Hypertensive Disease"
- Quality – individual entities which can be perceived or measured, such as "Color", "Length", "Blood Pressure" or "Sex"
- Region – mainly spatial, temporal and abstract "regions" for structuring attributes and "values" , like "Male" as a possible value for the quality "Sex"

We see DOLCE's philosophical background as especially important in domains like clinical medicine where concepts are often dependent on either social practice or individual human perception. Our choice for DOLCE also depends on the estimation that the proclaimed "cognitive bias" is more appropriate for representing a conceptual space covering several scientific domains with different views on the term disease, than the realism-based approach of the Basic Formal Ontology (BFO) [4]. While DOLCE captures ontological categories underlying natural language and human commonsense, BFO rather says that "each ontology … represents some partition of reality" and denies the validity of (man-made) concepts in any ontology intended to represent reality.

2.2. External Ontologies

Our ontology adapts several existing ontologies that are considered as standard resources in the biomedical domain. We do so by linking to them (i.e., we are annotating our classes with identifiers from the source ontology like the UMLS identifiers) or by directly integrating different parts of them (i.e., via OWL import or remodeling parts). (cf. Fig. 2)

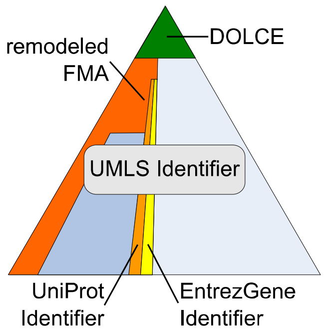

Figure 2. Links and inclusions of external terminologies and ontologies into the @neurIST ontology

The UMLS Metathesaurus [5] is one of the resources we have not directly integrated into our ontology but rather only linked to. It provides mainly taxonomic information about concepts, as well as synonyms and definitions for a large number of concepts. The concepts in our ontology were mapped to UMLS Concept Identifiers (CUIs) whenever possible by adding them as annotations to our concepts. The mapping to CUIs provides a preliminary classification of concepts according to the top level categories of the UMLS Semantic Network as well as a mapping to existing biomedical vocabularies containing adequate concepts. Biomolecular concepts are also directly linked to UniProt, EntrezGene and Gene Ontology by using their respective identifiers.

As example for the direct integration, the respective parts of the Foundational Model of Anatomy (FMA) [6] were integrated in our ontology by (re)modeling them in OWL DL along the lines of the DOLCE top-level. The FMA is a domain ontology representing the complete, canonical human body through explicit declarative anatomy knowledge. The intent is to assure through the FMA a consistent and standards-based representation of anatomical concepts in our ontology. But even though the FMA already has a very detailed anatomical model, the representation of neuroanatomical concepts was too weak for the needs of the project and therefore we expanded it with concepts deduced from literature search.

Subsequently, we combined the concepts from the integrated external sources with self-defined concepts by applying relations defined as OWL restrictions. Most of the basic relation types were already provided by the DOLCE ontology which defines e.g., mereological relations and participation relations. Therefore we only had to add domain specific relations when necessary. For this, we were following, as much as possible, the recommendations and definitions of the Relation Ontology that is part of the Open Biology Ontologies (OBO) Consortium [7].

2.3. Coverage Assessment

At the moment, the ontology contains about 3000 concepts and covers all of the projects' disciplines and sources. But it is actually problematic to assess the actual level of coverage: The inherent size and complexity of the ontology makes it hard for domain experts to properly judge whether it indeed contains all concepts necessary to fulfill certain application tasks. We have yet to come up with a straightforward solution to this problem but we see the mentioned web-based ontology browser as a first step into simplifying this task. After giving a brief introduction to the browser and an instruction on what to look for, domain experts were able to express their issues regarding the ontology, e.g., what concepts were missing or what concepts were modeled wrong. Still, the possibility to easily display the relations between concepts (and their logical implications) is a weak point and has to be improved.

3. Use Cases

The development of any ontology has to be driven by a comprehensive collection of use cases and requirements. The following lists the uses of the ontology in the project:

- Primarily, the ontology serves as a terminology in all project parts. It identifies the allowed terms and their respective meanings in both the clinical and experimental documentation as well as in the development of databases and user interfaces. The ontology gives textual and logical definitions for the concepts and is used as a standardizing instrument throughout the project where ambiguities are frequent (e.g., clinical terms). For adequate coverage, the ontology is further linked to separate lexical resources which are not part of the ontology proper. They provide preferred terms as well as synonyms and are (at least partially) multilingual.
- An important activity in the project is discovery of new knowledge. Therefore, our ontology provides several term lists (i.e., dictionaries) which can be employed in text mining and annotation application which are developed in cooperation with the respective project partners. Semantic analyses are supported though the relations specified between the different concepts.
- One of the most ambitious objectives of the project is the inclusion of distributed grid computing on a large scale. The tasks of the ontology in this scenario are the support of data mediation between heterogeneous databases on a semantic level and furthermore to serve as a knowledge repository for the binding of distributed services.

4. Usability

4.1. Context-independent vs. Context-dependent Knowledge

It is “nice” to have an ontology that fulfills all of the “principles of good ontology design” but to be really useable it also has to fulfill the specific needs of its users. To balance those two goals in our ontology, we decided to create two major parts, a “core”, context-independent part and one putting the concepts from the first one in specific contexts as required by project. In the first part, concepts are assigned strictly ontologically as subtypes of the DOLCE upper level ontology. The second part then specifically represents concepts in certain application-related contexts of human knowledge and makes it possible to extract subparts of the ontology according to user’s needs (e.g., for clinical, epidemiologic, or text mining purposes). The association to subtypes of the context-dependent part to the context-independent one is not performed by employing multiple inheritance but rather "on-the-fly" by classifying the context-dependent concepts according to the restrictions specified within the class definitions.

For example, in the context-dependent branch "Risk Factor for Intracranial Aneurysm State" all currently known proven and suspected risk factors are classified. Specifying the restrictions that lead to this classification as proven or suspected risk factors is based on an initial review of literature in the field of epidemiology. Currently, the ontology classifies 31 concepts as such risk factors (e.g., "Arteriovenous Malformation State" or "Polycystic Kidney Autosomal Dominant").

From a purely ontological standpoint it is obvious, that the introduction of such epistemological concepts is rather "un-ontological" [8]. But still, if the ontology is supposed to be of any practical use for the application tasks within the project then it is prohibitive not to include such concepts: They are obviously of tremendous importance in the daily work of our project partners and hence have to be included. But on the other hand, since we are aware of this "impurity", we provide a clear distinction between these epistemological concepts and the "truly ontological" ones by introducing the two distinct parts in our ontology based on this issue.

5. Conclusion

Following the objective of the @neurIST project to integrate heterogeneous data on a large spatial and temporal scale and originating from different sources, our ontology must incorporate conceptual spaces of such various domains as clinical medicine, molecular biology, imaging, physiological simulation and epidemiology. Based on the sound foundations of the DOLCE upper ontology we have represented about 3000 relevant concepts so far. The ontology is used at several parts of the project, as a controlled vocabulary for documentation and user interfaces, as basis for text-mining and annotation and also in software architecture for semantic data mediation. A questionnaire circulated among our project partners has shown that the ontology successfully fulfills all of the major requirements for the mentioned tasks.

References

[1] M. Boeker, H. Stenzhorn, K. Kumpf, P. Bijlenga, S. Schulz, S. Hanser, *The @neurIST Ontology of Intracranial Aneurysms: Providing Terminological Services for an Integrated IT Infrastructure*, In Proc. of the AMIA Annual Symposium, Chicago, USA, 2007.

[2] S. Hanser, J. Fluck, L. Furlong, C. Friedrich, M. Hofmann-Apitius, H. Stenzhorn, M. Boeker, *Knowledge Structuring and Retrieval for Intracranial Aneurysm Research.* In Proc. of the HealthGrid Conference, Chicago, USA, 2008.

[3] A. Gangemi, N. Guarino, N. Masolo, A. Oltramari, L. Schneider, *Sweetening ontologies with DOLCE.* In Proc. of the Internat. Conf. on Knowledge Engin. and Management (EKAW), Siguenza, Spain, 2002.

[4] P. Grenon, B. Smith, L. Goldberg, *Biodynamic Ontology: Applying BFO in the Biomedical Domain*, In Proc. of the Workshop on Medical Ontologies, Rome, Italy, 2003.

[5] O. Bodenreider, *The Unified Medical Language System (UMLS): integrating biomedical terminology*, Nucleic Acids Research 32 (2004), 267-270.

[6] C. Rosse C, J. Mejino, *A reference ontology for biomedical informatics: the Foundational Model of Anatomy*, Journal of Biomedical Informatics 36 (2003), 478-500.

[7] B. Smith, W. Ceusters, B. Klagges, J. Köhler, A. Kumar, J. Lomax, C. Mungall, F. Neuhaus, A. Rector, C. Rosse, *Relations in Biomedical Ontologies*, Genome Biology 6 (2005).

[8] S. Schulz, H. Stenzhorn, Ten Theses on Clinical Ontologies, In Proc. of the International Council on Medical and Care Compunetics Event (ICMCC 2007), Amsterdam, The Netherlands, 2007.