Ontological Modeling for Neurovascular Disease Studies: Issues in the Adoption of DL

Gianluca Colombo, Marco Antoniotti, Daniele Merico, Flavio De Paoli, and Giancarlo Mauri

Dipartimento di Informatica, Sistemistica e Comunicazione
Università degli Studi di Milano–Bicocca
Via Bicocca degli Arcimboldi 8, 20126 Milano, Italy
{gianluca.colombo, marco.antoniotti,}@disco.unimib.it

1 Introduction

We describe the ontological modeling issues encountered in the EU-funded NEUROWEB project. The aim of the project is to support association studies in the field of neurovascular medicine, integrating the different local repositories maintained by the clinical partners. Specifically association studies are carried out by searching statistical correlations between a feature and an *aggregate state* (i.e., *phenotype*), such as the occurrence of a complex/multi-factorial pathology. In medicine, and thus also in the specific neurovascular domain, the occurrence of a phenotype is implicitly asserted through the diagnostic process, an activity that is deeply rooted in the expert knowledge of the clinical community. Therefore, the major ontological commitment is to define phenotypes having a shared semantic, and connected to the data stored in local repositories. Since association studies are extremely sensitive to noise, a knowledge-intensive treatment is required.

2 A Critical Assessment of Existing Bio-Ontologies

Independent analyses of common-use, generalist biomedical ontologies (such as SNOMED-CT) have often revealed significant deficiencies. These are not primarily due to DL expressiveness limits, but rather to the adoption of too generic semantics for relations, which lead to inconsistencies in specific cases. For instance, Ceusters et al. [1] provide examples of erroneous subsumption relations in SNOMED-CT, as a consequence of using relationships that are too generic in their scope. This sort of representation defects is usually grounded on an insufficient conceptual modeling due to the lack of expert core knowledge and, in turn, it also leads to neglecting semantic issues being discussed in the formal logic community. Clinically-grounded phenotypes, the core entities to be modeled, are aggregate concepts, which can be deconstructed into fundamental building blocks. In analogy to the work done in the genomic field [2], we identify four conceptual components: (A) the *anatomical part* interested, (B) the *value*

observed, (C) the *device or method* used for the diagnostic exam, (D) the *etiology*, i.e. the complex of causes and risk factors concurring to the pathogenesis of the disorder. The modeling issues to be faced are two-fold, conceptual and formal: (1) clinical phenotypes are not explicitly described (e.g., in manuals); on the contrary, they are grounded on the core expert knowledge guiding the diagnostic process. (2) the aggregate nature of phenotypes requires a mereological treatment; therefore it is necessary to explicitly take into account the different formal semantics of part-of relations, as described both by foundational works in knowledge representation [3, 4] and in biomedicine ontological modeling [5, 6] fields. We argue that the availability of core expert knowledge is the enabling factor of an adequate use of formal relation semantics; as a matter of fact, in order to develop the conceptual model, we oriented Knowledge Acquisition to explicitly formulate the otherwise implicit mereological semantics assumed by the domain experts. Without core expert knowledge sources, the semantic specification of such relationships would be ungrounded. This strategy has lead to the development of a two-layered, inter-connected ontological framework [7]: the top level is a taxonomy of pathology types and subtypes; the bottom level enables to represent the building blocks underlying the pathological phenotypes, exploiting the different mereological relations already identified (anatomical part, value, diagnostic device, etiology). The formal language adopted is $\mathcal{SHIQ}$ providing the required expressive power for mereological relations, as discussed in [3]. The specific experience acquired through the NEUROWEB project suggests that semantic refinements, spawned by formal contributions, provide valuable guidance to improve the Knowledge Engineering task. In turn, an advancement in the field of Knowledge Representation requires an adequate Knowledge Engineering methodology, capable of supporting the adoption of a DL enriched of expressiveness for ontological modeling in specific domains.

References

1. Ceusters, W., Smith, B., Flanagan, J.: Ontology and medical terminology: Why description logics are not enough. In: Proceedings of TEPR 2003 - Towards an Electronic Patient Record, San Antonio, Texas (2003)
2. Bard, J.B.L., Rhee, S.Y.: Ontologies in biology: Design, applications and future challenges. Nature Reviews Genetics **6**(5) (2004/03) 213 222
3. Sattler, U.: Description logics for the representation of aggregated objects. In W.Horn, ed.: Proceedings of the 14th European Conference on Artificial Intelligence, IOS Press, Amsterdam (2000)
4. Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., Patel-Schneider, P.F., eds.: The description logic handbook: theory, implementation, and applications. Cambridge University Press, New York, NY, USA (2003)
5. Hahn, U., Schulz, S.: Towards a broad-coverage biomedical ontology based on description logics. In: Pacific Symposium on Biocomputing. (2003) 577–588
6. Bodenreider, O., Stevens, R.: Bio-ontologies: current trends and future directions. Brief Bioinform **7**(3) (2006) 256–274
7. Colombo, G., Merico, D., Antoniotti, M., Frisoni, G., De Paoli, F., Mauri, G.: An ontological modeling approach to neurovascular disease study: the neuroweb case. BMC Bioinformatics (2007) to appear.