An examination of OWL and the requirements of a large health care terminology

Kent Spackman

Department of Medical Informatics and Clinical Epidemiology
Oregon Health & Science University, Portland, Oregon, USA
spackman@ohsu.edu

Abstract. This paper presents a brief initial look at some of the possible benefits and barriers to using OWL as the language for the development, dissemination and implementation of terminological knowledge in the domain of health and health care. In particular, this assessment is made from the perspective of the author's role in the development of the Systematized Nomenclature of Medicine (SNOMED). To date, SNOMED has developed and adopted its own special-purpose syntax and formats for terminology development, exchange and distribution. Its representation language has limited expressivity yet is not expressible by any dialect of OWL 1.0. With the evolution to OWL 1.1, the barriers to using OWL for knowledge representation have been resolved. However, partly because of SNOMED's very large size, there remain barriers to adoption of OWL XML/RDF for SNOMED development, distribution or exchange purposes.

1 Introduction

The Systematized Nomenclature of Medicine, Clinical Terms (SNOMED CT) [1] is a work of clinical terminology with broad coverage of the domain of health care, and it has been selected as a national standard for use in electronic health applications in many countries, including the U.S., U.K., Canada, Australia, Denmark, and others. SNOMED was originally published in 1976, while SNOMED CT became available in 2002 as a major expansion resulting from the merger of SNOMED RT with the U.K.'s Clinical Terms version 3. A major distinguishing feature differentiating it from prior editions is the use of description logic (DL) to define and organize codes and terms [2].

Another major distinguishing feature of SNOMED is its size and complexity. With over 350,000 concept codes, each representing a different class, it is an order of magnitude larger than the next largest DL-based ontology of which we are aware. The size of the OWL XML/RDF form of SNOMED is approximately 248 MB, and this is just the DL representation without all the synonyms, mappings, subsets, and other special-purpose components of the terminology.

2 Knowledge Representation

As noted by Patel-Schneider [3], the design of OWL has been driven by three main streams of influence: the Semantic Web, description logics, and frame sys-

tems. To some degree these three streams can flow together, but in a more important sense they produce conflicting design criteria.

From the healthcare terminology perspective, the most useful stream of influences on OWL was the description logic stream. Long prior to the conception of the Semantic Web and the promulgation of related standards, the SNOMED effort had adopted description logic as the representation language for its formal definitions.

SNOMED CT was developed with a description logic which includes conjunction ($\sqcap$), existential role restrictions ($\exists R.C$), role hierarchy axioms ($r \sqsubseteq s$), and class definitions of the form $A \sqsubseteq C$ with A a class name and C an arbitrary class description. In addition, it makes use of what have been termed "right identities", a restricted form of property chain inclusion axioms of the form $r \circ s \sqsubseteq r$, where the circle denotes role composition. This relatively limited set of language constructs is nevertheless sufficient to represent a very wide range of meanings.

Another important property of this limited language, which can be called $\mathcal{EL}\mathcal{H}$ with right identities, is that its computational complexity remains polynomial[4], and there exist classifiers that can handle the task of classifying the entire terminology[5].

Without a solid DL foundation, the Semantic Web would have remained largely irrelevant to health care terminology standardization. Even so, the initial version of OWL was developed without taking adequate account of features of DL that had already been used in both the GALEN [6] and SNOMED [7] efforts. The development of OWL 1.1 eliminated one of the most significant barriers to use of OWL for SNOMED, since it permits the identification of tractable sublanguages capable of handling the size and complexity of SNOMED[8]. Although adding property chain inclusion axioms was reportedly the most difficult step in developing OWL 1.1 [9], this was essential. Without it, adoption of OWL by the SNOMED community would have required awkward workarounds with their attendant complications and complexities – effectively killing movement in that direction. With it, we have a clear path to using OWL 1.1 for further development and integration with other biomedical ontologies.

3 Development and Exchange

Beginning in 1995-96, SNOMED's developers adopted a distributed co-operative model for modeling the DL definitions of its meanings [10]. This development model was driven by several factors, including the geographical distribution of health care professionals available to work on curation of the terminology, as well as a recognition of the value of distributed development.

In order to successfully handle asynchronous modeling by geographically dispersed editors, software was developed that permitted each editor to submit "change sets" which could be imported into a central configuration management environment and analyzed for conflicts. The format of these change sets has evolved over time, and SNOMED has recently defined an published an open source specification for an XML interchange format [11].

As mentioned above, scale is a major issue for SNOMED. This comes into play when considering the most common tools available for editing and development of OWL-based work. All open source editing tools that I have tested run into memory problems when dealing with a terminology as large as SNOMED. While some of these barriers will be (or perhaps by the time of this conference have been) removed as a result of special programming effort, it is probable that only a few open source editing tools will receive this level of refinement and enhancement. It is unclear whether most ontologies and terminologies for the semantic web are small-scale because that is the right size, or merely because the tools and resources to build truly useful large terminologies are not available. In health care, at least, it is clear that we need a very large and well-integrated terminology, such as SNOMED.

To the extent that OWL assumes that Semantic Web resources will be developed independently and that publication is effectively a static expression of these resources on the Web, then there is no need for standards that permit broad and distributed development, critique, curation, correction, enhancement, or modification. On the other hand, our experience with SNOMED suggests strongly that distributed development is crucial, and therefore it would be important to consider support of this need by OWL standards. Also based on our experience, support for change sets, at least, would not seem to be a very difficult thing to add to the standard. SNOMED's change set documents are available under the Apache 2.0 license and could be used as a starting point if desired.

4 Distribution

In order to be useful, SNOMED has to be integrated into application programs that are in use in the health care environment. Of course, virtually none of the commercial applications in wide use in health care today are built to take account of the Semantic Web. As a result, each vendor or software supplier or health care institution has a slightly different need for terminology standards, and incorporates those standards into their application software in different ways.

There is an existing distribution format for SNOMED, based on a set of files of UTF-8 characters suitable for loading into database tables. In order to provide application developers with a stable specification, this distribution format was submitted for American National Standards Institute (ANSI) standardization[12]. This structure has served its purpose for several years. However, as the knowledge representation evolves and there is a need for more expressiveness, the current format has become limiting. As a result, a special-purpose XML format has been developed and published for comment[13]. Once again, it would seem that it would be worthwhile to examine the representation capabilities of OWL 1.1 as a means for representing at least the DL-based components of SNOMED. Other components, such as change tracking, mapping, subsetting, and so forth may not be as readily represented in OWL. However, these issues facing terminologies in health care are probably not unique to this one domain, and it would be worth examining the extent to which an enhanced OWL standard might

contribute to improved use of terminologies across several industries, thereby increasing the chances of the Semantic Web vision being realized.

5 Conclusion

We have discussed several issues that have arisen in the process of examining the utility of OWL and its associated products for the purposes of supporting ongoing efforts to formalize representation of terminology for health and health care. The move to OWL 1.1 is seen as a very positive move from a knowledge representation standpoint. Support for large scale and distributed development of ontologies is seen as a need that is as yet unmet.

References

1. SNOMED *Clinical Terms*. Northfield, IL: College of American Pathologists, 2007.
2. F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, and P. F. Patel-Schneider, editors. *The Description Logic Handbook. Theory, Implementation, and Applications*. Cambridge, U.K.: Cambridge University Press, 2003.
3. P. F. Patel-Schneider. What is OWL (and why should I care)? Ninth International Conference on the Principles of Knowledge Representation and Reasoning. Whistler, Canada, June 2004.
4. F. Baader, S. Brandt, and C. Lutz. Pushing the $\mathcal{EL}$ envelope. In *Proc. of the Nineteenth Int. Joint Conf. on Artificial Intelligence (IJCAI-05)*, Edinburgh, UK, 2005. Morgan-Kaufmann Publishers.
5. F. Baader, C. Lutz, and B. Suntisrivaraporn. CEL—a polynomial-time reasoner for life science ontologies. In U. Furbach and N. Shankar, editors, *Proc. of the 3rd Int. Joint Conf. on Automated Reasoning (IJCAR'06)*, volume 4130 of *Lecture Notes in Artificial Intelligence*, pages 287–291. Springer-Verlag, 2006.
6. A. L. Rector, S. K. Bechhofer, C. A. Goble, I. Horrocks, W. A. Nolan and W. D. Solomon. The GRAIL Concept Modelling Language for Medical Terminology, *Artificial Intelligence in Medicine*, 9:139-171, 1997.
7. K. A. Spackman. Managing clinical terminology hierarchies using algorithmic calculation of subsumption: Experience with SNOMED-RT. 2000. OHSU Technical Report.
8. B. C. Grau, I. Horrocks, B. Parsia, P. Patel-Schneider and U. Sattler. Next Steps for OWL. Proceedings of the Second OWL Experiences and Directions Workshop (OWLED) 2006.
9. M. Horridge, D. Tsarkov, and T. Redmond. Supporting early adoption of OWL 1.1 with Protege-OWL and FaCT++. Proceedings of the Second OWL Experiences and Directions Workshop (OWLED) 2006.
10. K. E. Campbell, S. P. Cohn, et. al. Galapagos: Computer-based support for evolution of a convergent medical terminology. Proc AMIA Annu Fall Symp 1996, pp. 269-273.
11. SNOMED Interchange Format Specification. College of American Pathologists, 2006. http://www.snomed.org/snomedct/interchange_format.html.
12. Healthcare Terminology Structure: ANSI Standard. College of American Pathologists, 2003.
13. SNOMED XML Schema Specification. College of American Pathologists, 2006. http://www.snomed.org/snomedct/xml_schema.html.