Ontology Archaeology: Mining a Decade of Effort on the Suggested Upper Merged Ontology

Adam Pease, Chris Benzmueller[1]

1 EXTENDED ABSTRACT

The Suggested Upper Merged Ontology [5] is a large ontology defined in first-order logic with some higher-order extensions [1]. The project began in the year 2000. Each version has been released open source and publicly from the start, which provides a unique record of the construction of a formal ontology. While initially just an upper ontology, it now encompasses a wide variety of domains, and some recent work has involved semi-automatically merging large factbases with the fully axiomatized hand-built content [3]. SUMO has been mapped by hand to all of English WordNet [6] and several other languages (Elkateb et al 2006; Borra et al 2010; Pease& Fellbaum, 2010) and is used in natural language understanding tasks (Pease & Li, 2010). SUMO is supported by tools for ontology development (Pease, 2003) and inference (Trac et al 2007) and used in the yearly CASC theorem proving competition (Pease et al 2010).

Before discussing some of the specifics of the development history, we must note that the historical data is still incomplete. A change in employment of the first author resulted in some data on the development of the domain ontologies being lost prior to mid-2004, when Articulate Software was founded. This accounts for the large non-linearity in the graph shown in Figure 1 at that time.

For the first two years, development was confined to the original upper level effort: the Suggested Upper Merged Ontology. The first step in the project involved seeding the ontology with the contents of all general-purpose, and formally defined axioms that we could find in open, academic work and merging those theories into a common structure. This resulted in about 5700 lines of SUO-KIF code that was released May 11, 2001. After that point, the majority of axioms were developed by the SUMO project team, although many more small theories and additions were provided by others over the entire history. All contributions are credited in the CVS logs for the ontologies [13]

In June 2002 a large US government project provided support for creating a host of new domain ontologies that would extend SUMO. These ontologies covered information about economy, finance, government, geography and many other topics necessary for encoding facts about the world's geo-political situation. It is these domain ontologies which would form the largest addition of domain content to SUMO for some time, and which is responsible in the graph for the first major jump in content size.

Early in 2003 we began a comprehensive project to add content that would structure and define concepts more general than the new domain ontologies, yet more specific than SUMO itself. This would become known as the MId-Level Ontology (MILO). The methodology for creating MILO was an outgrowth of our work in mapping SUMO to WordNet (Niles&Pease, 2003). After the original mapping was complete, it was clear that a vast number of linguistic terms (or what in WordNet are called synsets) lacked a mapping to an equivalent formally defined concept in SUMO. We therefore set out to create roughly equivalent formal concepts for every synset that appeared at least 3 times in the WordNet semantic concordance, or SemCor (Miller et al 1993). Since the completion of that effort roughly a year later, MILO has continued to be an active site for development of new content as additional domains are covered, and new intermediate-level content is required to structure and adequately define terms that apply to many domains. More recent significant new domains that can be seen increasing the size of the overall theory were many military concepts added in 2006 and digital media concepts in 2010.

While SUMO has always viewed manual creation of formal axioms as the primary effort, the breadth and formality of content has recently enabled databases of lightly structured information to be integrated automatically, and to take advantage of the definitions that SUMO provides. Simple factbases alone provide little opportunities for significant inferences, but combined with thousands of formal rules, many new and useful conclusions become possible through automated inference. The mappings to Mondial, portions of DBPedia, the Open Biomedical Ontologies, and YAGO have been significant. The content of these databases dwarfs the hand-created content in SUMO, as shown in Figure 2, which shows the same time period as Figure 1, but with, in order from the bottom, DBPedia, OBO and YAGO added.

SUMO			
Total Terms	relations	Total Axioms	Rules
1173	353	4741	932
MILO			
Total Terms	relations	Total Axioms	Rules
1662	159	5116	1183
Domain ontologies			
Total Terms	relations	Total Axioms	Rules
17312	708	77974	2041
Total for all ontologies			
Total Terms	relations	Total Axioms	Rules
20147	1220	87831	4156

Table 1: SUMO term and axiom statistics

[1] Articulate Software, Angwin, CA, USA. Email: {apease, cbenzmueller}@articulatesoftware.com.

Lines of code are a useful approximation of the progress of ontology development, but insufficient. The Sigma browser provides statistics of how many terms are in the knowledge base, and how many of those terms are relations. It also shows how many axioms there are, and how many of those axioms are "if..then" rules. Those numbers are shown in Table 1. Note that we do not include totals for YAGO, which dwarfs the hand-coded content, and totals some 16,425,285 axioms.

REFERENCES

[1] Benzmueller, C., and Pease, A., (2010). Progress in Automating Higher-Order Ontology Reasoning, to appear in Practical Aspects of Automated Reasoning workshop of CADE 2010..

[2] Borra, A., Pease, A., Roxas, R., Dita, S., (2010). Introducing Filipino WordNet, in Principles, Construction and Application of Multilingual Wordnets: Proc. of the 5th Global WordNet Conf., Bhattacharyya, et al, editors, ISBN:978-81-8487-083-1.

[3] de Melo, G., Suchanek, F., and Pease, A., (2008). Integrating YAGO into the Suggested Upper Merged Ontology. In Proceedings of the 20th IEEE Int'l Conf. on Tools with AI (ICTAI 2008).

[4] Elkateb, S., Black, W., Rodriguez, H, Alkhalifa, M., Vossen, P., Pease, A. and Fellbaum, C., (2006). Building a WordNet for Arabic, in Proceedings of LREC 2006.

[5] Miller, G. A., Leacock, C., Tengi, R., and Bunker, R. T. 1993. A semantic concordance. In Proceedings of the Workshop on Human Language Technology (Princeton, New Jersey, March 21 - 24, 1993). Morristown, NJ, 303-308.

[6] Niles, I., & Pease, A., (2001), Toward a Standard Upper Ontology, in Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001), C. Welty and B. Smith, eds..

[7] Niles, I., and Pease, A., (2003). Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology, Proc. of the IEEE Int'l Conf. on Information and Know. Eng. pp 412-416.

[8] Pease, A., (2003). The Sigma Ontology Development Environment, in Working Notes of the IJCAI-2003 Workshop on Ontology and Distributed Systems. Volume 71 of CEUR Workshop Proc. series.

[9] Pease, A., and Fellbaum, C., (2010) Formal Ontology as Interlingua: The SUMO and WordNet Linking Project and GlobalWordNet, In: Huang, C. R. et al (eds.) Ontologies and Lexical Resources. Cambridge: Cambridge University Press, ISBN-13: 9780521886598.

[10] Pease, A., and Li, J. (2010) Controlled English to Logic Translation. In Theory and Applications of Ontology, ed. Roberto Poli, Michael Healy, and Achilles Kameas, Springer, ISBN: 978-90-481-8846-8.

[11] Pease, A., Sutcliffe, G., Siegel, N., and Trac, S., (2010). Large Theory Reasoning with SUMO at CASC, AI Communications, Volume 23, Number 2-3 / 2010, Special issue on Practical Aspects of Automated Reasoning, IOS Press, ISSN 0921-7126, pp 137-144.

[12] Trac, S., Sutcliffe, G., and Pease, A., (2008) Integration of the TPTPWorld into SigmaKEE. Proceedings of IJCAR '08 Workshop on Practical Aspects of Automated Reasoning (PAAR-2008). Volume 373 of the CEUR Workshop Proceedings.

[13]CVS logs for SUMO and its domain ontologies: http://www.ontologyportal.org/SUMOhistory/ , http://sigmakee.cvs.sourceforge.net/sigmakee/KBs/

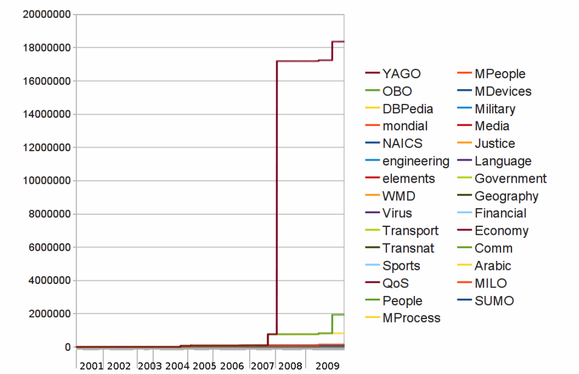

Figure 2: Automatic additions of database content December 2000 to present

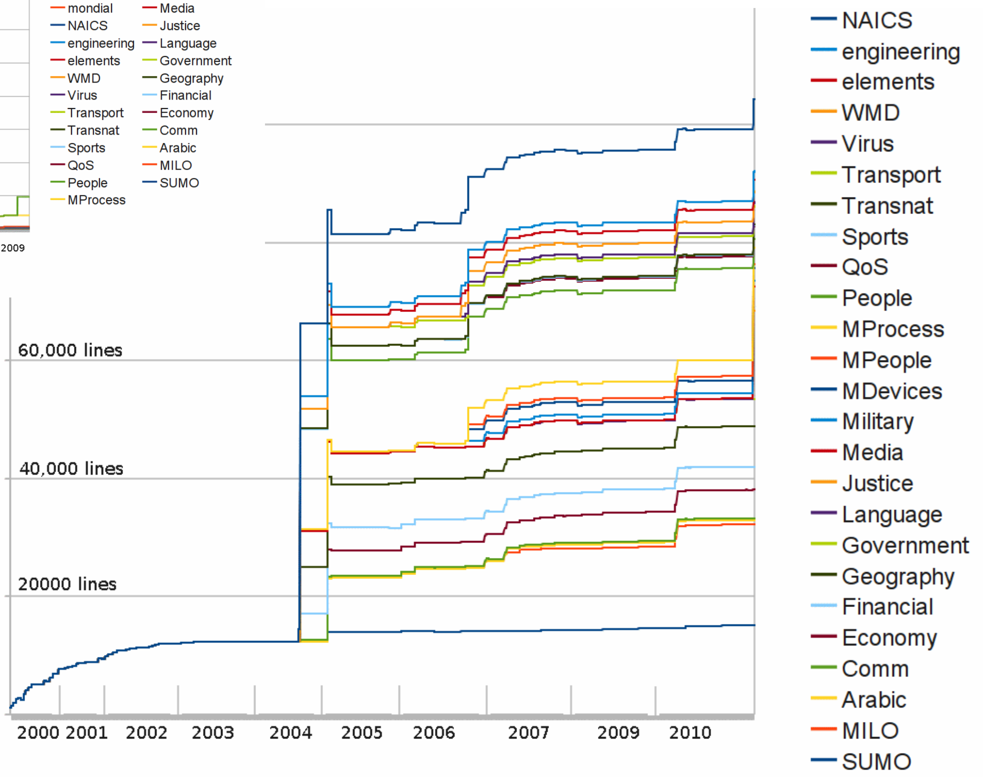

Figure 1: SUMO Family of Ontologies: Lines of Code Totals from December 2000 to Present