Wrestling with SUMO and bio-ontologies

To the editor:

As researchers in bioinformatics turn increasingly to ontological tools and resources to help them to solve problems of information integration, data annotation, natural language processing and automated reasoning, the need for proven, useful ontologies in biomedical research will continue to grow exponentially. We are delighted to see *Nature Biotechnology* taking up this important issue, with the publication of a commentary on the MGED (Microarray Gene Expression Data Society) ontology in the September issue by Soldatova and King (*Nat. Biotechnol.* **23**, 1095–1098, 2005).

Soldatova and King make the important point that ontology builders need to be aware of emerging standards and best practices. When they trace the origins of modern work on ontology to Aristotle, however, it is important to remember that philosophers have been debating Aristotle's categories continuously for more than two millennia. The number of ways in which we can slice up the world into categories is in practice unlimited, and it is not always obvious which distinctions one needs to make, for the purposes of the MGED ontology, or for any other purpose. Moreover, finding problems in virtually any extant ontology is a trivial exercise. Soldatova and King recommend the use of the Suggested Upper Merged Ontology (SUMO) as an emerging standard with the capacity to bring coherence to ontologies in the biomedical domain. Unfortunately, SUMO in its current form embodies no well-defined criteria to determine which classes should be properly included within its scope, with unfortunate consequences for its overall integrity and usability for purposes of ontology integration in the life sciences. Should an upper ontology designed for purposes of robust integration of biomedical ontologies include classes such as *Monkey* or *BodyCovering*? Or odd disjunctive classes, such as *FruitOrVegetable*?

It is self-evident that methods to assist the empirical evaluation of both ontology content and structure are urgently required. As Soldatova and King themselves acknowledge, "the engineering of ontologies is still a relatively new research field."

Much of the most influential ontology work in biomedicine has been stimulated by the pressing needs of bench biologists themselves in managing burgeoning quantities of data. As a consequence, many of the ontologies developed thus far are somewhat unprincipled in comparison to what we now know can be achieved. Today, however, we have reached the point where an increasing number of biomedical scientists are recognizing the importance of learning about standards of good practice in ontology development and of adhering to those standards whenever possible.[1]

The newly created National Center for Biomedical Ontology[2], formed under the US National Institutes of Health (NIH) roadmap, is a direct acknowledgment of this need. We are principal participants of this center and have a particular interest in improving the quality of all ontologies developed for use in biomedicine. The center will be attempting, through systematic outreach activity and through testing and dissemination of good ontology practices, to aid biomedical investigators in the construction of ontologies that adhere to proven conventions and knowledge-representation formalisms[3]. We will conduct workshops designed to promote collaboration among different groups of ontology developers and to assist biomedical researchers in developing and applying ontologies precisely tailored to their needs. We believe that the establishment of the center will offer an opportunity to enhance consistency and clarity in biomedical ontologies and to increase the prospects for their interoperability, for example, through the use of common, well-defined relationships along the lines applied to all new entries in the Open Biomedical Ontologies library[4,5].

The central role of good ontologies in biomedical informatics is unquestioned. What we need now is research to establish how best to achieve our broader goals through the formalization and integration of biomedical knowledge.

Mark A. Musen[1], Suzanna Lewis[2] & Barry Smith[3]

[1]Stanford University, Stanford Medical Informatics, 251 Campus Drive, Suite X-215, Stanford, California 94305-5479, USA, [2]Lawrence Berkeley National Laboratory, Berkeley Drosophila Genome Project, 1 Cyclotron Drive, Mailstop 64-121, Berkeley, California 94720, USA and [3]University at Buffalo, Department of Philosophy, 126 Park Hall, Buffalo, New York 14260, USA.
e-mail: musen@stanford.edu

1. Wang, X., Gorlitsky, R. & Almeida, J.S. *Nat. Biotechnol.* **23**, 1099–1103 (2005).
2. http://www.bioontology.org
3. http://protege.stanford.edu/publications/ontology_development/ontology101.html.
4. Smith, B. *et al. Genome Biol.* **6**, R46 (2005). Published online 28 April 2005 doi:10.1186/gb-2005-6-5-r46.
5. http://obo.sourceforge.net/relationship/

To the editor:

As researchers involved in the development of the MGED Ontology (MO) and other bio-ontologies, we were pleased to see *Nature Biotechnology* foster dialogue on the challenges in building robust and optimal ontologies for biomedical research in a commentary by Soldatova and King published in the September issue (*Nat. Biotechnol.* **23**, 1095–1098, 2005). However, we wish to address several misleading and inaccurate descriptions of the development and use of the MO, and comment on the motivation behind and constraints inherent in the development of such bio-ontologies.

Ontology development in biology has a good track record for addressing real and immediate needs for describing and classifying biological and experimental data—genes, proteins, experiments, tissues, treatments, functions. It has done this in a