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

manner that affords straightforward, reliable and increasingly automated access to both data and the methods by which they were generated, and is helping to meet the time-critical demands of current research. The MO is one of the early examples of an ontology that was developed by a community in that spirit, specifically to describe microarray experiments. In their commentary, Soldatova and King show no awareness of the challenges posed by rapidly evolving, high-throughput genomic technologies that are generating data at an exponentially increasing rate. Controlled terminology as provided through an ontology was urgently required for the description of microarray experiments; biologists did not expect emerging ontologies to be ideally engineered, and we would be the first to state that the MO contains compromises. In the absence of such controlled terminology, much of the data being generated today would not be available in a form readily amenable to comparative analysis, either now or in the future.

Fundamental inaccuracies in this critique appear to reflect a lack of understanding of the MO's nature. For example, the authors state that the MO "has been promoted as an international standard." The MAGE object model is an international standard for exchange of microarray data. The MO was developed to provide terminology required to support the annotation of microarray experiments in areas where the MAGE Object Model does not provide descriptors. Furthermore, the authors state that the Minimum Information About a Microarray Experiment (or MIAME) specification is at the root of some of the design compromises in the MO. This is a puzzling assertion; surely they meant to refer to the MAGE Object Model, from which some of the classes in the MO were created.

The authors suggest rebuilding the MO. But the benefits of such refactoring need to be weighed against the necessity of maintaining a stable real-world resource that is widely applied in software implementations. This is in fact the reason why there is a 'core' component to the MO deliberately held stable—a design decision that seems to be misunderstood by the authors. Additionally, the Functional Genomics Ontology (or FuGO; http://fugo.sf.net) project is now underway to collaboratively develop controlled terminology for transcriptomics, proteomics and metabolomics, as used in a range of biologically defined domains. The decomposition and reuse of the MO will be part of the process of building FuGO, allowing the MO itself to remain stable, continuing to serve its established user base.

The authors suggest the building of ontologies that are (i) purpose independent and (ii) compliant with a standard upper ontology. These 'key rules for bio-ontology development' represent unrealistic expectations of the current state of play as the adoption of an upper ontology is predicated on its acceptance by user communities who must decide whether or not an upper ontology is fit for its intended purpose. There are many 'standard' upper level ontologies. Such ontologies are themselves subject to major philosophical debate. In addition, a valid demonstration of their utility in ontology building is by no means universally accepted, even within the computer science community.

Finally, we address perhaps our greatest concern with respect to this commentary: the article centers on the "isolation of bio-ontologies from the larger world of ontologies" and asserts the need for interaction between MO developers and other ontologists. It is therefore ironic, and of notable concern, that the authors of this commentary have never communicated their analysis, or suggestions for improvements to any of the extended family of MO developers, that includes ontologists in an advisory capacity. This sort of collaborative resource development is not a competitive exercise. We have ongoing interactions with a number of leading ontologists, and we encourage other researchers to contact us (http://mged.sourceforge.net/ontologies/) to share their views and help ensure that good ontologies, which benefit the bioscience community, are developed as effectively and efficiently as possible.

Christian Stoeckert[1], Catherine Ball[2], Alvis Brazma[3], Ryan Brinkman[4], Helen Causton[5], Liju Fan[6], Jennifer Fostel[7], Gilberto Fragoso[8], Mervi Heiskanen[8], Frank Holstege[9], Norman Morrison[10], Helen Parkinson[3], John Quackenbush[11], Philippe Rocca-Serra[3], Susanna-Assunta Sansone[3], Ugis Sarkans[3], Gavin Sherlock[2], Robert Stevens[10], Chris Taylor[3], Ronald Taylor[12], Patricia Whetzel[13] & Joseph White[11]

[1]University of Pennsylvania School of Medicine, Genetics, 1415 Blockley Hall, 423 Guardian Dr., Philadelphia, Pennsylvania 19104, USA, [2]Stanford University, Department of Biochemistry, Stanford University School of Medicine, 269 Campus Drive, Stanford, California 94305-5307,USA. [3]European Bioinformatics Institute, EMBL Outstation, Hinxton, Cambridge CB10 1SD, UK, [4]BC Cancer Research Centre, 675 West 10th Avenue, Vancouver, BC V5Z 1L3, Canada, [5]CSC/IC Microarray Centre, Imperial College, Hammersmith Campus, DuCane Road, London W12 ONN, UK, [6]KEVRIC, an IMC Company, 8484 Georgia Ave., Suite 550, Silver Spring, Maryland 20910, USA, [7]National Institute of Environmental Health Sciences, 111 T.W. Alexander Drive, Research Triangle Park, North Carolina 27709, USA, [8]Center for Bioinformatics and Computational Biology, National Institute of General Medical Sciences, Building 45, Room 2AS55K, 9000 Rockville Pike, Bethesda, Maryland 20892, USA, [9]Genomics Laboratory, University Medical Center Utrecht, STR 3.223, Universiteitsweg 100, 3584 CG Utrecht, the Netherlands, [10]University of Manchester, School of Computer Sciences, Oxford Road, Manchester M13 9PL, UK, [11]Department of Biostatistics, Dana-Farber Cancer Institute, Mayer 232, 44 Binney Street, Boston, Massachusetts 02115, USA, [12]Computational BioSciences Group, Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington 99352, USA and [13]University of Pennsylvania, 1415 Blockley Hall, 423 Guardian. Drive, Philadelphia, Pennsylvania 19104, USA.
e-mail: stoeckrt@pcbi.upenn.edu

To the editor:

We read with interest the commentary by Soldatova and King in the September issue (*Nat. Biotechnol.* **23**, 1095–1098, 2005). We applaud the editors for providing a forum for discussion of issues that will improve our ability to create ontologies for the purpose of integrating data from many communities. Many of the points made about unwise use, appropriate naming and unclear definitions in the MGED ontology were very useful.

We also feel that some points in the review warrant comment. First, the underlying assumption of the paper is flawed. It is misleading to state that by critiquing the MGED ontology one can critique many or even all ontologies in biology. It would be a mistake for readers to conclude that currently all existing bio-ontologies fail to follow international standards for ontology design and description. The title should instead have been 'Is MGED a good ontology?'

Second, the critique is misleading in some of its arguments: of the ten issues asserted as needing detailed study, two of these (issues nine and ten) reflect only design choices and do not affect the reasoning efficiency of the ontology.

Third, Soldatova and King fail to consider some lessons that have been learned from the literature of ontology evaluation (e.g., as part of evaluating the distinction between 'is-a' and 'part-of' relations, tools such as OntoClean[1] can be used for taxonomy cleaning; these tools are currently used with WordNet for upper level taxonomy).