

Are the current ontologies in biology good ontologies?

Larisa N Soldatova & Ross D King

The failure of many bio-ontologies to follow international standards for ontology design and description is hampering their application and threatens to restrict their future use.

"Progress will still need to be driven by the logic of genetics and by further increases in abstraction."
Edward B. Lewis, Nobel lecture

The research histories of both biology and ontologies originate with the philosopher Aristotle. For 2,400 years, the two subjects have taken separate paths. Now, with the rise of bioinformatics, they are reunited in the 'hot' research topic of biological ontologies. So why are biological ontologies important? As more and more biological data are stored on computers, the problem of efficient retrieval and analysis of these data becomes the most important scientific bottleneck, and the problem is particularly acute in biology because biological data are notorious for their complex form and semantics. Ontologies can help because they embody the abstract knowledge required for data integration and analysis. The utility of ontologies has been clearly demonstrated in several biological domains (e.g., Gene Ontology¹). However, within biology, the enthusiasm for ontologies has been accompanied by a general lack of awareness of what exactly ontologies are and how to use them. This lack of awareness is reflected in the fact that many, perhaps all, 'bio-ontologies' fail to follow international standards for ontology design and description. This failure is important because it places a serious restriction on their applicability to knowledge sharing, reuse and inference.

Larisa N. Soldatova and Ross D. King are at the Computer Science Department, The University of Wales, Penglais, Aberystwyth, Ceredigion, SY23 3DB, UK.
e-mail: lss@aber.ac.uk

In this article, we analyze the current state of application of ontologies in biology, try to reveal the reasons for the existing difficulties, recommend a possible solution to the current problems and describe prospects and future challenges for the application of ontological engineering in biological domains. We focus on the example of the ontology developed by the Microarray Gene Expression Society (MGED) for describing microarray experiments². We have chosen this example not because it is particularly bad, but because it exhibits the typical problems of biological ontologies, and because MGED has been promoted as an international standard. We wholeheartedly agree that such standards are extremely valuable, but we also believe that their promoters are particularly obliged to get them right, otherwise their whole advantage is negated.

Ontological engineering and biology

The engineering of ontologies is still a relatively new research field. There as yet does not exist a well-developed theory and technology for ontology construction, as there is for bridge construction, for example. This means that many of the steps in designing an ontology remain manual and a kind of 'art'³. An ontology consists of four main components: concepts (or classes), a hierarchical structure (*is-a* relations, a backbone of an ontology), relations (other than *is-a* relations), and axioms. Riichiro Mizoguchi⁴, one of the leaders of ontological engineering, lists fundamental ideas of a class, an instance and *is-a* relations in ontology:

Intrinsic property. The intrinsic property of a thing is a property that is essential to the thing, which loses its identity when the property changes.

The ontological definition of a class. X is a class if and only if (iff) each element x of X satisfies the intrinsic property of X. If and only if (iff) this definition holds then the relation $\langle x \text{ instance-of } X \rangle$ is true.

Is-a relation. $\langle \text{class } A \text{ is-a class } B \rangle$ relation holds between classes if and only if (iff) every instance of the class A is also an instance of the class B.

The Institute of Electronics and Electrical Engineering (IEEE; Washington, DC, USA) Ontology Working Group is developing an upper ontology standard. This upper ontology is intended as an anchor for all other ontologies. It will contain the most general categorizations of knowledge, a unified language for ontology representation and axioms for ontology constraints⁵. The idea is that if other domain ontologies follow this standard, such as those in biology, it will enable the sharing and reuse of knowledge across different compliant domain ontologies. It will also provide a prototype for domain ontologies, and a guide for ontology construction. Unfortunately, few, if any, ontologies in biology are compliant with an IEEE standard upper ontology or follow their standard recommendations for ontology design. The intersection between the ontologies listed at <http://suo.ieee.org/SUO/Ontology-refs.html> and those in the open biological ontologies repository (<http://obo.sourceforge.net/cgi-bin/table.cgi>) is the empty set. Workers in bio-ontologies seem content to plough their own furrow and ignore the wider ontological world.

The isolation of bio-ontologies from the larger world of ontologies is especially important in the growing fields of biological knowledge discovery and computational inference⁶. To enable correct logical inference, ontologies



Figure 1 MGED Ontology (a fragment). The MGED ontology is divided into two parts: the core ontology which describes the most essential concepts about microarray experiments, and the extended ontology. This fragment shows the detailed concept hierarchy for <ExperimentPackage> and <ExperimentDesignType> concepts.

must not only share their concepts, but also unify types of relations between them and used axioms. Otherwise, the full potential of ontologies will not be realized. These problems are illustrated by the components of the MGED Ontology.

Analysis of the MGED Ontology

The MGED Ontology was developed by MGED to provide descriptors required for MAGE v.1 (MicroArray and Gene Expression) documents. It is aimed to be the basis of the MIAME (Minimum Information About a Microarray Experiment) standard for capturing core information about microarray experiments and provides a conceptual structure for microarray experiment descriptions and annotation (**Fig. 1**). The MGED Ontology is one of the first attempts to formalize the description of experiments in biology. We enthusiastically agree that such formalizations are extremely important for organizing and executing experiments, as well as storing, and sharing of the experimental results.

How well then does the MGED Ontology meet its stated aims? If we analyze the actual annotations of real experiments deposited in microarray databases (e.g., the Array Express database at European Bioinformatics Institute⁷), we find that MGED does not presently contain nearly enough terms to describe actual microarray experiments. For example, in the list of experiment objects

of the experiment 'E-CAGE-11,' there are 37 concepts, of which only 9 belong to the MGED Ontology. Specifically, the ontology of microarray experiment does not include such concepts as <Array>, <Annotation>, <PhysicalBioAssaySource>, <Identifier>, <Name>, <Description> of an experiment. The ontology is not able to clearly describe the details of protocols and the steps required in an experiment.

This limitation probably lies behind the MAGE working group's confession: "The boundaries between MIAME concepts...and the MGED ontology (that try to define and structure the MIAME concepts) is neither well defined nor easy to understand"⁸. What are the reasons for MGED incompleteness and incomprehensibility? Below we list some of the problems with the MGED constraint design.

The MGED Ontology (version 1.1.9, September 3, 2004. Since the submission of our manuscript, a new version of the MGED ontology MO 1.2.0 has been released. Unfortunately, none of the problems we describe have been solved in this new version.) consists of 226 'classes,' 644 'individuals' and 109 'properties.' There are no definitions of the terms 'classes,' 'individuals' or 'properties,' which itself is a serious error of omission. However, after studying this use, we believe that it may be taken from the ontology language DAML (the DARPA agent markup language⁹). Given this, analysis of MGED constraints reveals the following problems:

1. MGED appears to contain ontologically incorrectly determined classes. All members of a defined class should share at least one intrinsic property (see above). A typical case of not doing this is that MGED defines the class <BioAssayPackage> as "MAGE package for bioassay" with subclasses <BioAssay>, <DeriveBioAssayType> and <ImageFormat>. There appears to be no single intrinsic property of all the instances of these classes; for example, what do the JPEG format and the mean_and_coefficient_of_variation share in common? Other examples are <DescriptionPackage> and <DesignElementPackage>.
2. It is unwise to use the same name at different levels of abstraction. The class <Individual>, defined as "identifier or name of the individual organism from which the biomaterial was derived," is a subclass of <BioMaterial Characteristics>. This is confusing as 'individual' is a type of meta level object of the MGED ontology (see above).
3. Some concepts appear to be incorrectly named. For example, <BioMolecularAnnotation> is defined as "BioMolecularAnnotation experiment design types are those which are designed to investigate functions, processes, locations and identity at the molecular level, e.g., binding site identification, genotyping." From this definition, we learn that the class is a kind of experiment design type, but the name suggests that it is a kind of annotation.
4. Some concepts appear to be inappropriately named. It is important in an ontology that appropriate names are chosen for classes, as one of the main functions of an ontology is to make the contained knowledge explicit to both human and computer reasoners. However, MGED names such as <...Package> may be acceptable to object-orientated programmers, but are not appropriate in an ontology aimed at biologists.
5. MGED appears to have several incorrect definitions of classes. For example, the class <ProtocolType> has three subconcepts (types of a protocol); however, according to its definition, <DataTransformationProtocolType> is not a protocol, but a physical process: "The process by which derived BioAssays are created from measuredBioAssays and/or derivedBioAssays." The IEEE Ontology Working Group's upper ontologies all clearly distinguish physical processes and other abstract concepts. Therefore,

there is a contradiction and MGED cannot be used with ontologies from domains outside of biology.

6. Some definitions in MGED are also unclear. For example, the class <Experiment> is “the complete set of bioassays (hybridizations) and their descriptions performed as an experiment for a common purpose. Here we take experiment to mean an observational or perturbing study. An experiment will be often equivalent to a publication.” These are actually three separate definitions; do all three need to be true? We also fail to see how an experiment is equivalent to a publication.

7. MGED confuses concepts and procedures. To connect with other ontologies, the class <OntologyEntry> is used. However, it is really not a class, but a procedure for connecting to other ontologies. (Note, the standard language for coding ontologies, OWL, is especially designed to provide an easy access to concepts of other ontologies.)

8. MGED does not properly distinguish between a class and an individual. Why are <absolute>, <ORF> and <RNA> considered individuals and not classes? Some concepts, such as <Atmosphere>, are considered to be both a class and an individual. This may be necessary in some circumstances, but it seriously complicates the reasoning process.

9. The structure of MGED Ontology allows multiple inheritances of properties. For example, the individual <chromosome> is a member of two parent classes: <TheoreticalBioSequenceType> and <PhysicalBioSequenceType>, or an “abstraction used for annotation” and at the same time a “biological sequence that can be physically placed on an array.” The nature of an abstract concept and a physical object are traditionally considered different in philosophy, and for good reasons. For example, with abstract chromosomes it is possible to duplicate them at will and reason about infinite sets of them (for phylogenetic reasoning); the same is not true for physical chromosomes. A directed acyclic graph lattice structure, like that of the MGED structure, is acceptable by the IEEE Ontology Working Group⁵; however, it puts a number of severe restrictions on the use of the ontology for logical inference, and therefore should only be used if it is absolutely necessary.

10. In MGED, the relationships between concepts are not well defined. Instead of using

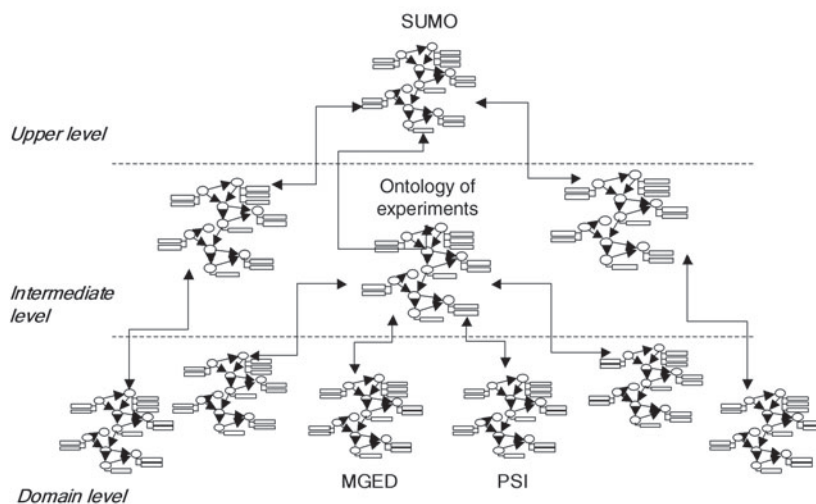


Figure 2 An ontology repository. Upper level ontologies, such as SUMO, describe the most general concepts about the world. Domain ontologies, such as MGED and PSI, formalize experimental domain concepts. A domain-independent ontology of experiments at an intermediate level would provide a general prototype for description of domain-dependent experiments, and ensure full compliance with the ontology repository

part-of and *attribute-of* relations, the MGED ontology defines 109 properties of classes and individuals, such as *has_ID*, *has_factor_value_ontology_entry*, *has_measurement*, *has_property_set*, *has_action*. In total, 104 *have - X* types of properties from the list of 109 properties. This unnecessarily complicates the ontology, decreases its comprehensiveness and makes reasoning less efficient.

11. There is also an unclear distinction in MGED between the use of *is-a* and *part-of* relations. For example, <Experiment> and <ExperimentDesign> are subclasses of the class <ExperimentPackage>, but you would naturally expect that <ExperimentDesign> to be a *part-of* the concept <Experiment>, which it is not.

The extended MGED ontology is still ‘under construction.’ However, there seems to be no clear strategy guiding what to place in the core ontology, and what to put into the extended ontology aimed to add further associations to MAGE v.1. For example, the concept <ProtocolVariation> defined as “the effects of different protocols or changes in protocols on experimental results are studied” is from the extended ontology; in contrast, the concept <MethodologicalFactorCategory>, defined as “the effects on results of changing protocols, hardware, software, or people performing the experiments,” is from the core ontology. Yet according to their definitions, the first concept is a subconcept of the second one. It is

also unclear why <BioAssayData> (defined as “files including images generated from one or more BioAssays”) belongs in the extended ontology, whereas the core ontology contains <Measurement> (defined as “measured values and units”) and <MaterialType> (defined as “examples are population of an organism, organism, organism part, cell, etc.”) along with definitions of a protein, a cell, a virus and a whole organism. To add to this confusion, the US National Cancer Institute Distributed Terminology Server (DTS) browser for MGED gives a different MGED version (<http://nciterns.nci.nih.gov/NCIBrowser/>).

The MGED core ontology stores information about dimensions (Armstrong, liter, Kelvin, mole), formats (GIF, JPEG or TIFF), types of publication (a book, a journal article or online resources) with mistakes similar to those mentioned above: a <book> is an individual, not a class; the concept <BibliographicReference> does not have a direct connection with <PublicationType>; and the concept <title> is used, but is not determined, either as a class or as an individual. Note, that much of this knowledge does not strictly belong to the domain of microarray experiments. Instead, it is common for any type of experiment and more generally to other domains outside of science. Therefore, they should be kept separately in a higher level ontology (see below). The Suggested Upper Merged Ontology (SUMO) already has formalized knowledge about dimensional units. A better strategy would be a connection to SUMO with the reuse of its concepts, perhaps

adding missing ones, but not duplicating them with the introduction of mistakes.

Towards an ontology of scientific experiments

Can we preserve the wealth of useful information about microarray experiments in MGED, while making it compliant with ontological standards? Unfortunately, given its many problems and misconceptions, we are forced to conclude that the best approach is a clean start with the reuse of substantial parts, rather than an attempt to reengineer the existing ontology. The reason for this is that in an ontology, it is generally possible to change some concepts, names and definitions, and add axioms, but it not easy to change the structure. We are forced to conclude that it would be better to reuse a part of the concepts describing domain knowledge, shift domain-independent classes to the upper level of abstraction and completely change the structure. It has already been announced that the MGED Ontology will be a prototype for modeling, capturing and annotating proteomics experimental data¹⁰. It is therefore urgent that the problems with MGED are solved as soon as possible, before they are repeated in new domains with the subsequent waste of scientific and financial resources.

If we were to start with a clean sheet, then the standard engineering approach would be to store generic knowledge about scientific experiments separately from knowledge about specific domains (Fig. 2). Experimental goals, methods, requirements, experimental restrictions, and rules for experiment design are found not only in microarray experiments, but also in most experiments in biology, and more generally in any science¹. We therefore propose the development of an Ontology of Experiments that would lie intermediate between upper ontologies and specific scientific domain ontologies. The advantage of this abstraction is that generic knowledge is held in only one place, ensuring consistency and clean updating. Many domains of study use similar instruments and material as do scientific experiments. Such knowledge, as well as information on dimension units, data types and types of bibliographic references should

be represented at the level of upper ontology. Likewise, descriptions of experiment objects and subjects such as a human, an animal, a plant, a robot, belong to scientific domain ontologies such as MGED. The division of knowledge into corresponding levels and the integration of the ontological knowledge for retrieval and inference will enhance the comprehensiveness and functionality of bioinformatics systems, optimize their design and help to avoid many mistakes made in their construction.

Conclusions

From the beginning of the application of ontologies to bioinformatics, there have been debates about how best to form ontologies⁶: whether to take time to ensure that the foundations are sound or to ignore the niceties of ontological logic. In practice, almost all biological ontologies have been generated in a quick (and arguably dirty) manner and have generally ignored ontological logic. The case for doing it was (and is) that the ontologies are needed as soon as possible and they are primarily designed to provide working biologists with a common vocabulary for standard annotation purposes; after all, human biologists have the intelligence to get around any contradictions. This is a strong argument, and there can be no denying the success of such ontologies as Gene Ontology. However, the approach is not compatible with the increasing use of computational reasoning in biology and its dependence on ontological data. Expert biologists may be able to deal with poorly designed and inconsistent ontologies, but this not currently possible for computer programs that do machine learning or text mining. As such programs are set to dominate the analysis and retrieval of biological data, we argue that biological ontologies should be designed for their needs as well.

In this context, new biological ontologies should avoid the types of mistakes shown by MGED and follow instead unified standards and design methodologies. The majority of bio-ontologies were built to provide a common vocabulary and for a standard annotation; however, ontologies have far greater potential and could open up whole new pos-

sibilities for biological research. The construction of bio-ontologies as simple taxonomies restricts their current and future application. The formation of a set of integrated ontologies at different levels of representations would significantly increase interoperability between domain data and knowledge, and enable new intelligent bioinformatics applications.

The key rules for bio-ontology development are the following:

- Explicitly list the principles of an ontology's design, its constraints, along with definitions and axioms.
- Provide compliance with a standard upper ontology (SUO) developed by IEEE P1600.1. The ontology society has no accepted standard yet, SUMO, OpenCyc or a lattice of multiple upper ontologies, but any of them might be a reasonable guideline for construction of bio-ontologies.
- Keep separately domain-dependent and domain-independent knowledge, as well as declarative and procedural knowledge, to provide efficient sharing and reuse knowledge.
- Build ontologies so that they are purpose-independent and therefore are 'future-proof'.

We argue that ongoing efforts to develop bio-ontologies need to adopt these rules. Anything less will likely prevent them from fulfilling their great potential in biology.

1. <http://www.geneontology.org>
2. <http://mged.sourceforge.net/ontologies/MGEDontology.php>
3. Sowa, J. Knowledge Representation. Logical, Philosophical, and Computational Foundations (Brooks/Cole, New York, 2000).
4. Mizoguchi, R. *New Generation Computing* 22/2, 193–220 (2004).
5. <http://suo.ieee.org/SUO/SUMO/index.html>
6. Schulze-Kremer, S. *Prac. Int. Conf. Intell. Syst. Mol. Biol.* 5, 272–275 (1997).
7. <http://www.ebi.ac.uk/arrayexpress/>
8. http://www.mged.org/Workgroups/MIAME/miame_image-om.html
9. <http://www.daml.org>
10. <http://psidev.sourceforge.net/gps/>