Bio-ontologies as Tools for Integration in Biology

Sabina Leonelli
Department of Economic History
London School of Economics
London, UK
sabinaleo@hotmail.com

Bio-ontologies are a relatively recent achievement of the bioinformatic effort toward an efficient organization and distribution of biological data. They provide a structured, controlled vocabulary through which data—especially those gathered through sequencing and genomics, but increasingly also those resulting from other types of research—can be classified in a form that can be stored in and retrieved from online databases. As Walter Gilbert predicted, back in 1991, sequencing technologies have pushed biologists to rethink their approach to sharing and using data (Gilbert 1991). The opportunity to use new digital resources, especially the software and infrastructure developed within information technology as part of the Semantic Web, has strengthened the ongoing emphasis on data-driven research.[1] Bio-ontologies play a central role in this process, by providing a common classification system to be used in any database collecting data on one or more model organisms. Tools such as the Gene Ontology and the Plant Ontology are becoming prominent standards facilitating the display of data within open-access databases and thus their circulation across research contexts.[2]

The increasing popularity of bio-ontologies needs to be understood against the background of three defining features of contemporary biology. The first is the vast fragmentation into local epistemic cultures, each of them studying different organisms through a variety of experimental and conceptual toolkits. This has led to the proliferation of vocabularies to describe the same phenomena (either attributing different meanings to the same term or using different terms to refer to the same thing), a linguistic and methodological pluralism resulting from the finely tuned relationship between each group and the objects it studies. Indeed, researchers shape their terminology to fit their tacit knowledge and their conceptual understanding of the organisms they study. There is therefore much epistemic richness in linguistic diversity, yet that same diversity is also an obstacle to communication across research contexts.

The second factor is the enormous amount of heterogeneous experimental data produced by these communities. Assessing the biological significance of those data involves finding ways to assemble, order, and use them to inform new research—that is, to make them reusable across research contexts. The third factor is the common pursuit of an integrated understanding of biological processes informed by genomic knowledge (often associated to the term "system," even if system biology is but one of the approaches fostering knowledge integration). Developing tools to bridge across local research cultures is crucial to satisfying this need.

The idea of bio-ontologies originated in the early 1990s from model organism communities, and specifically from the curators of some of the first databases specializing in one organism. These curators realized that the storage and distribution of the ever-growing amounts of data on the fruit fly, the mouse, and yeast were achievable only through a common ordering structure and terminology. Especially since completion of the first sequencing projects, the sheer size of datasets and variability in methods of disclosure (publications, public repositories, patents) have made collecting available data on any one organism a very difficult task. The issue is tied to the practical problem of how to store these data and make them searchable. Whatever the structure, standards, and terminology employed for this purpose, the system chosen must be computer-readable so as to support automated data analysis. It must also be intelligible and accessible to as many fields and expertises as possible, rather than based on the needs and issues characterizing one community. What is needed, in short, is an independent classification system that can potentially be adopted by any model organism database and can incorporate requests from a variety of communities.

Reflecting these needs, the Gene Ontology was proposed as a classification system that aims to enhance the availability and usability of data gathered on several different organisms (currently numbering 12, including *Homo sapiens*) across

March 21, 2008; revised and accepted March 26, 2008
Biological Theory 3(1) 2008, 7–11.

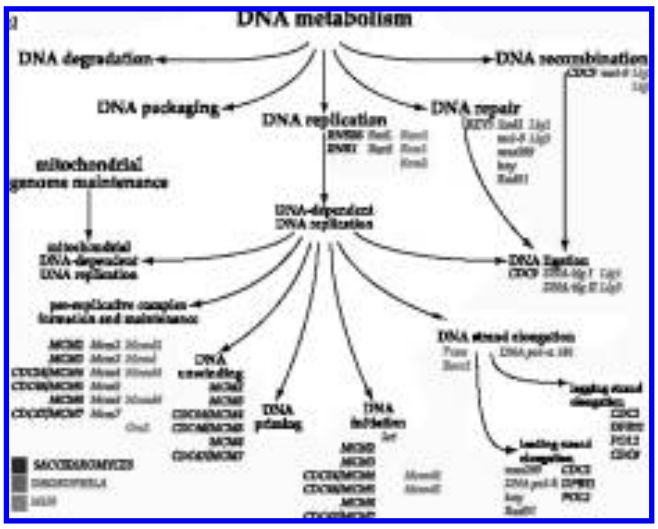

Figure 1.
A schematic representation of terms and relations used in the branch of the Gene Ontology devoted to "biological processes." A similar hierarchical structure is used to classify terms pertaining to "cellular components" and "molecular functions."

research contexts. The Gene Ontology Consortium includes representatives from all the main databases (and thus their respective communities of users) wishing to make use of the Gene Ontology. Through regular meetings and communication, the expertise and requests of each participating database are brought to the attention of bio-ontology curators, thus helping to shape their work. At the same time, the Gene Ontology facilitates the retrieval of gene, gene product, and sequence annotations by anyone interested in using the data for their own research purposes. The case of the Gene Ontology exemplifies how bio-ontologies might provide a practical solution to the seemingly elusive quest for biological integration: they promise to deliver a unique gateway to the results of biological research without challenging its sophisticated fragmentation into subcommunities. Whether they succeed in this goal depends on their ability to respond to the diverse needs and competences of their users. In what follows, I sketch how bio-ontologies achieve such flexibility and reflect on whether this approach differs from the long-lived ideal of a unified scientific language.

Developing a Controlled Vocabulary

Building a bio-ontology involves two main steps: the elaboration of a "controlled vocabulary" whose terms are given a precise definition and are related to each other in some specified way,[3] and the association of these terms (via machine-readable symbols called "unique identifiers") with datasets gathered through experiment, a process referred to as gene annotation and carried out in collaboration with databases using the bio-ontology (Bard and Rhee 2004: 213).[4] These two phases are always intertwined in practice, as curators need to select terms whose meaning and relations to each other make biological sense and at the same time prove appropriate for the classification of existing datasets.

I view the constant balancing of terminological and data-driven concerns as one of the most important epistemic features of bio-ontologies. Bio-ontologies are tools for displaying and analyzing the results of experimental work. Accordingly, the primary aim of their vocabulary is to classify data. As for all other classification systems, however, the first step toward developing a controlled vocabulary is to choose relevant classificatory categories. In the case of bio-ontologies, this involves deciding which phenomena are to be associated with datasets, so that researchers can easily check which experimental findings are associated with the specific phenomena they are investigating. Curators do not base this decision on their own ideas about what counts as a phenomenon in biology, but rather consult practicing biologists and scientific publications. The content of bio-ontologies is thus shaped by the current research landscape rather than by curators' ideas about what entities and processes should be of relevance to biological research.

Bio-ontologies are not the product of a vision or an interpretation of what counts as biological knowledge, nor are they meant as a representation of biological ontology in the strictly philosophical sense: they do not attempt to describe all that exists independently of human intervention, concentrating rather on the entities and processes currently under investigation. In other words, the ontology underlying the bio-ontological vocabulary concerns the objects of biological practice. By grounding classification on data-producing practices, this approach recognizes that what biologists now consider to be a biological entity or process depends at least in part on current interests, knowledge, and theoretical perspectives and might change in the course of future research. As often emphasized by curators, "existing [bio-ontological] terms are augmented, refined and reorganized as the current state of biological knowledge advances" (Gene Ontology Consortium 2006: D322). This emphasis on dynamic development illustrates the close link between the evolving content of biological knowledge, as produced and shaped by research in the lab, and the way in which such knowledge is depicted in bio-ontologies for the purposes of classifying and circulating data. As highlighted by an influential review of standards in bioinformatics, "introducing standards should not be a goal in itself, but should help biologists to solve problems" (Brazma et al. 2006: 601). In other words, the standards should serve research rather than impose constraints on it.

There is a significant danger of imposing constraints on research when implementing a classification system such as a bio-ontology. This is because the terminology used in a bio-ontology needs to be coherent and economic: there can be no more and no less than one term for each entity or process of interest, so as to avoid repetitions, redundancies, and confusion.

Figure 2.
The Gene Ontology as it appears online. Users can click on each term to see which other terms it is associated with, and in which way, as well as to check which datasets are categorized under each term.

Prima facie, this requirement is in direct conflict with the need to serve communities that often use different languages to refer to the same phenomena. Curators have, however, developed strategies to accommodate this issue. For a start, curators select the term to be associated with each phenomenon of interest. Eventual ambiguities in the meaning of the term are cleared through the formulation of a definition of each term (Baclawski and Niu 2006: 35).[5] Once terms are chosen and defined, curators examine research contexts in which a different term is given the same definition or where the same term is defined in a different way.[6] To accommodate the former option, curators create a system of synonyms associated with each chosen term. For instance, the term "virion" is defined by the Gene Ontology as "the complete fully infectious extracellular virus particle." Given that some biologists use the term "complete virus particle" to fit this same definition, this second term is listed in the database as a synonym of "virion." Users looking for "complete virus particle" are thus able to retrieve data relevant to the phenomenon of interest, even if it is officially labeled "virion."

Curators use another strategy for cases of substantial scientific disagreement on how a specific term should be defined: The qualifier "*sensu*" allows them to generate subterms to match the different definitions assigned to the same term within different communities. This is an especially efficient strategy when it comes to dealing with species-specific definitions of terms. For example, the term "cell wall" is relabeled "cell wall (*sensu Bacteria*)," which is defined as peptidoglycan-based, and "cell wall (*sensu Fungi*)," which contains chitin and beta-glucan.

As long as curators are aware of differences in the use of terms across communities, the differences can be registered and assimilated in the bio-ontology so that users from all communities are able to query the corresponding database. Acquiring a good overview of the different uses of terms across all relevant research contexts is the main problem encountered by bio-ontology curators. Curators have created several ways of getting feedback from practicing biologists, including so-called "content meetings" (where experts in various fields are called on to comment on specific terms), online surveys, and dialogue at conferences and workshops. These methods are not as yet efficient in eliciting feedback, as researchers working at the bench often have neither the time nor the motivation to critically assess the structure and definitions used in bio-ontologies. Providing feedback requires some acquaintance with the processes through which bio-ontologies are developed and put to use, including familiarity with basic information technology and with the software and structures used in constructing the bio-ontology. Experimenters often resent the idea of learning so much bioinformatics: what they want is to use bio-ontologies to further their own research. In the eyes of many experimenters, the production of a reliable bio-ontology is the job of curators (their "service" to the community) and users should trust the curators' judgment rather than spend time in questioning it.

The current division of labor between bio-ontology curators and practicing biologists might change if, as now seems likely, researchers are asked to submit data to databases whenever they wish to publish a scientific paper.[7] This experience might encourage direct involvement by experimenters in the development and use of bio-ontologies. For example, if you wish to submit data to any database participating in the Gene Ontology Consortium, you need to know how to classify the data under Gene Ontology terms. An agreement of this kind has recently been implemented between *Plant Physiology*, a prominent journal in plant biology, and The Arabidopsis Information Resource [TAIR], the main database for data on *Arabidopsis thaliana* and a prominent user of the Gene Ontology (Ort and Grennan 2008). All researchers submitting a paper containing *Arabidopsis* data are asked to add those data to the TAIR database, which makes use of the Gene Ontology. It remains to be seen whether and how this will change plant biologists' perception of bio-ontologies.

A Unifying Language?

Bio-ontologies such as the Gene Ontology define a set of related terms that accurately and uniquely classify phenomena under investigation for the purpose of disseminating available data on those phenomena. In this sense, they provide a unifying language for biology that can be used by each participating community regardless of its specific epistemic culture, interests, and location. The strength of this approach is that, at least in theory, such unifying language exists only

for the purpose of exchanging data across research contexts. Individual researchers do not need to adopt bio-ontological terms in their daily practice and in their communications with their peers: they need only to know how to translate their own preferred terminology into those terms (for instance, through synonyms and *sensu* qualifications), so as to be able to consult databases and retrieve data of interest to their research.[8]

In this role of mediators between local contexts, bio-ontologies constitute powerful tools for integration. They have the potential to facilitate collaboration across communities of biologists, who can use them to disseminate their data and see what data are produced by other researchers. At least in principle, this function can be fulfilled without forcing members of different research contexts to unify their goals, methods, or—most importantly—their knowledge and vision of biological processes. Seen in this light, the role of bio-ontologies in biology may come close to what Otto Neurath once called a "neutral language"—a language facilitating social coordination by providing the "possibility of discussing actions in a common language and with common arguments, which may afterwards be translated into the phrases which sound familiar to people accustomed to the language of some creed or party" (Neurath 1944: 30). What makes bio-ontologies especially interesting is that their vocabulary is used not to describe actions or arguments but rather to classify data—thus leaving matters of interpretation in the hands of database users.

There are, however, reasons to question whether bio-ontologies will actually be able to fulfil this role in the long term. One concerns the degree to which data interpretation is left to users rather than curators. Curators need to make several judgments when choosing bio-ontological terms, defining them, and associating them with data. Not all of these judgments mirror precisely the approach to data taken by experimentalists. This is for the simple reason that curators are engaged in a different endeavor—that of assessing which terms are most likely to be recognized and understood by researchers across communities. On the basis of their understanding of each relevant field, curators choose which terms should be used as official bio-ontological language and which are to be treated as synonyms (if they are at all considered). They also decide which datasets are associated with which terms—a decision based mostly on interpretations already published and corroborated in the literature, but still requiring familiarity with each of the research contexts involved, so as to represent their claims as faithfully as possible. Of course, these types of interventions are unavoidable and bioinformaticians are specially trained to perform them. Yet the question remains: Are bio-ontologies truly neutral in their portrayal of research?

Another possible hurdle is the insufficient interaction between bio-ontology curators and users. As mentioned above, it is not clear whether users of bio-ontologies will simply accept the terms and definitions given by curators or critically engage with their content. The degree of accessibility of a bio-ontology depends on whether it incorporates terms that are recognized and used in all relevant communities. If little input is garnered from potential users, there is a risk of excluding existing research communities from accessing the databases. As a consequence, some biologists are likely to avoid using bio-ontologies because they do not trust their classification system. Indeed, there is a danger that, like many classification systems before them, bio-ontologies are shaped by (and impact on) power relations among research communities. The adoption of certain terms rather than others may favor specific research groups, for instance, by adding visibility to their work. This is especially true if, as seems likely, bio-ontologies become an indispensable part of journal submission procedures.

Many bio-ontology curators are well aware of these issues and are committed to both fostering outreach to potential users and maintaining the dynamic character of their system. This commitment is one of the principal reasons why several prominent bio-ontologies are associated under the institutional umbrella of the Open Biomedical Ontology, or "OBO Foundry," which acts as a platform through which curators can exchange material and coordinate the content of the various bio-ontologies.[9] Bio-ontologies should continue to be regarded as a useful service under constant improvement, rather than as a language to be trusted regardless of whether it fits researchers' own agendas. Excessive reliance on these classification systems might threaten the very feature that makes them so useful: the representation of actual research practices rather than curators' own ideas of what those should be.

Notes

1. Researchers in genomics are increasingly resorting to data analysis through databases (such as correlations and "random walks") to determine new research directions.

2. For information on the Gene Ontology, one of the best developed to date, see Ashburner et al. (2000) and Gene Ontology Consortium (2006); on the Plant Ontology, see Ilic et al. (2007).

3. For example, the Gene Ontology Consortium has chosen to classify the relations holding among objects into two categories: "is_a" and "part_of." The first category denotes relations of identity, as in "the nuclear membrane is a membrane"; the second category denotes mereological relations, such as "the membrane is part of the cell." Occasionally, a third category, "develops_from," is used to signal dependence relations, as in "protein develops from amino acids." In other ontologies (for instance, the ones employed to gather data about phenotypes), the available categories of relations are more numerous and complex; they include, for example, relations signaling measurement ("measured_as") or belonging ("of_a").

4. I am overlooking a third important step in the development of bio-ontologies: the elaboration of appropriate software and infrastructure supporting bio-ontologies. See Leonelli (2008) and Gene Ontology Consortium (2006, 2007) for a discussion of this issue.

5. For instance, in the Gene Ontology the term "nucleus" is defined as "a membrane-bounded organelle of eukaryotic cells in which chromosomes are housed and replicated. In most cells, the nucleus contains all of the cell's chromosomes except the organellar chromosomes, and is the site

of RNA synthesis and processing. In some species, or in specialized cell types, RNA metabolism or DNA replication may be absent (see http://www.geneontology.org).

6. The terms and approaches used to investigate a specific phenomenon vary even across groups that possess the same expertise but are studying different organisms. The Gene Ontology Consortium has the merit of having been one of the first institutionalized platforms encouraging collaboration and comparative work across model organism communities.

7. This move is primarily meant to provide an incentive for biologists to disclose their data. The free distribution of data is a necessary requirement for tools such as databases and bio-ontologies to work efficiently, but biologists still show a reluctance to submit their data to public repositories. Reasons for this range from concerns about the time and effort involved in making such donations to constraints posed by sponsors, fear of data appropriation by competing groups, and security issues.

8. Note that curators' avoidance of a priori ontological assumptions in classifying biological data does not make bio-ontologies into a unique labeling system in biology. Arguably, other classification systems have been constructed in the same pragmatic vein—most notably, as recently argued by Staffan Mueller-Wille (2007), Linnaeus' taxonomy. What might well mark a difference with other classification systems is the reliance of bio-ontologies on cutting-edge digital infrastructures, which makes them widely accessible and extremely flexible to user needs.

9. See the Open Biomedical Ontologies Consortium (http://www.obofoundry.org).

Acknowledgments

This research was funded by the Leverhulme Trust (grant no. F/07004/Z) and the ESRC as part of the project "How Well Do 'Facts' Travel?" based at the Department of Economic History, London School of Economics. Thanks to my research group, particularly Mary Morgan, and to audiences at Egenesis and ISHPSSB 2007 for helpful discussion.

References

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene Ontology: Tool for the unification of biology. Nature Genetics 25: 25–29.

Baclawski K, Niu T (2006) Ontologies for Bioinformatics. Cambridge, MA: MIT Press.

Bard JBL, Rhee SY (2004) Ontologies in biology: Design, applications and future challenges. Nature Reviews Genetics 5: 213–222.

Brazma A, Krestyaninova M, Sarkans U (2006) Standards for systems biology. Nature Reviews Genetics 7: 593–605.

Gene Ontology Consortium (2006) The Gene Ontology (GO) project in 2006. Nucleic Acids Research 34 (Database issue): D322–D326.

Gene Ontology Consortium (2007) The Gene Ontology (GO) project in 2008. Nucleic Acids Research Advance Access: 1–5.

Gilbert W (1991) Towards a paradigm shift in biology. Nature 349: 99.

Ilic K, Kellogg EA, Jaiswal P, Zapata F, Stevens PF, Vincent LP, Avraham A, Reiser L, Pujar A, Sachs MM, Whitman NT, McCouch SR, Schaeffer ML, Ware DH, Stein LD, Rhee SY (2007) The Plant Structure Ontology, a unified vocabulary of anatomy and morphology of a flowering plant. Plant Physiology 143: 587–599.

Leonelli S (2008) Circulating evidence across research contexts: The locality of data and claims in model organism research. Working Papers on the Nature of Evidence: How Well Do "Facts" Travel, No. 25/08. http://www.lse.ac.uk/collections/economicHistory/pdf/FACTSPDF/2508Leonelli.pdf

Mueller-Wille S (2007) Collection and collation: Theory and practice of Linnaean botany. Studies in History and Philosophy of Biological and Biomedical Sciences 38: 541–562.

Neurath O (1944) Ways of life in a world community. The London Quarterly of World Affairs (July): 29–32.

Ort DR, Grennan AK (2008) Editorial: Plant physiology and TAIR partnership. Plant Physiology 146: 1022–1023.