Basic considerations for improving interoperability between ontology-based biological information systems

Robert Hoehndorf

Ontologies are used in biology for the description of multiple kinds of entities. Large ontologies provide categories and relations for the basic features found in databases of model organisms. They serve as the basic means to integrate the data that is generated and interpreted by multiple heterogeneous groups and stored in distributed biological databases throughout the world. The use of a common vocabulary and common formal descriptions of the vocabulary's terms permit the comparison, retrieval and analysis of the data stored in these databases. The ontologies that are used for this purpose are primarily isolated, single-domain ontologies that have little or no interconnections specified among them. Ontology communities such as the Open Biomedical Ontologies (OBO) and the OBO Foundry establish guidelines to maintain quality and reusability of ontologies, and to facilitate interoperability between ontologies that are included in these projects.

I identify several facets of interoperability between ontology-based information systems in biology which are not currently addressed satisfactorily. First, the knowledge representation languages used to represent ontologies must be sufficiently rich to express the distinctions made by the ontology designers, and required by the applications of the ontology. Second, the basic categories of the biological ontologies must be analyzed and integrated within a common conceptual framework to permit information to flow between the ontologies. Finally, to let information flow between domain ontologies, the acquisition of additional knowledge from domain experts is required.

Most biological ontologies are represented in the OBO Flatfile Format and the Web Ontology Language (OWL). I propose extensions to both forms of representing biological ontologies. The semantics of the OBO Flatfile Format is not explicit, and the current proposals for a semantics of the OBO Flatfile Format do not coincide with the way it is used in many ontologies, in particular in statements that use negation. Therefore, I propose a more flexible semantics through a translation to OWL. The decidable version of OWL is equivalent to an expressive description logic. However, it is based on classical logics and exhibits the property of *monotonicity*. When combining ontologies, it is beneficial to consider alternative, non-classical logics that permit *nonmonotonic* inferences. I propose a method for integrating biological ontologies which are formalized

either in the OBO Flatfile Format or OWL using a default logic.

Core ontologies provide an ontological foundation for domain ontologies by extending top-level ontologies with domain-specific axioms. They can be used to integrate domain ontologies and as a starting point for the development of new ontologies within a domain. I introduce the biological core ontology *GFO-Bio*. GFO-Bio is implemented in OWL and first order logic, and is accompanied by axioms in default logic. I include several elaborated modules in GFO, such as a module for biological functions, disposition or biological sequences. Additionally, I illustrate how GFO-Bio can be used to integrate biological domain ontologies and facilitate information flow among them.

To integrate biological domain ontologies using GFO-Bio or any other top-level or core ontology, additional knowledge about the interrelations between domain categories must be acquired from domain experts. Due to the large number of categories in these ontologies, such an effort is time-consuming and expensive. Methods and software applications that permit a large number of domain experts to collaborate on this task would enable the rapid and cheap acquisition of ontological knowledge. For this purpose, I introduce the BOWiki, an ontology-based semantic wiki, and a social tagging system. In addition, I suggest several novel methods for automatically extracting data and knowledge from natural language texts. Automated extraction of biological knowledge can provide an alternative to manual curation of ontologies and their annotations, or serve as a starting point for manual efforts of knowledge acquisition.

The primary focus of this work is the development and discussion of novel methods for improving interoperability between biological domain ontologies. These are classified in three major categories, and the relations between them are analyzed. I show how their application leads to improved interoperability and increased usability of the ontologies.