An Ontology-Based System for Querying Life in a Post-Taxonomic Age
(Collaborative Research; NSF Advances in Biological Informatics, Innovation track)

Public Abstract

Improving our understanding of life, whether the biology of individual species such as our own, or the mechanisms and processes governing biodiversity at large, critically depends on integrating, querying, and aggregating biological data from many different organisms. To this day, the most fundamental and common way to accomplish this relies on organism names, making these one of the pillars of querying and managing our biological knowledge and data. However, the traditionally used names for organisms and groups of organisms, which are based in Linnaean nomenclature, suffer from two major limitations to their usefulness when it comes to integrating and communicating data. First, because they are simple text-strings, the meaning intended by those who coin a name and those who apply it is inaccessible to machines. As a result, exactly which organisms a name is or is not meant to include is often ambiguous, and names are therefore often applied inconsistently. Second, there are many groups of organisms that do not yet and may never have a Linnaean name, but for which molecular or macroscopic characteristics have been discovered that constitute valuable biological knowledge. This project aims to address these issues by generating a mechanism, called *phyloreferencing*, that allows referring to any group of organisms of shared evolutionary descent by a machine-interpretable definition of the unique pattern of descent that distinguishes the group from all others. This project builds on more than a decade's worth of theoretical and applied research into phylogenetic taxonomy. With the recent synthesis and continuous update of a universal phylogenetic Tree of Life, phyloreferences will have immediate and broad practical applications for communicating, integrating, and querying biological data across the Tree of Life. In contrast to authoritative nomenclatural naming, the goal for phyloreferences is that users can construct them instantly for any group of shared evolutionary descent for which they wish to communicate discoveries. In addition to developing the phyloreferencing standard and supporting tools, this project also aims to create classroom-ready teaching materials and curricula for training biology students in the theoretical underpinnings and practical applications of phylogenetic nomenclature.

The phyloreferencing mechanism that this project will develop uses standards and tools developed for the Web, specifically the Web Ontology Language (OWL), ontologies, and machine reasoning. Specifically, this work aims to develop phyloreferences as OWL class expressions consisting of sufficient and necessary conditions, which machine reasoners can use to identify subclasses and class instances. Ontology and reasoning technologies have already shown their power for biological knowledge integration and discovery and are increasingly being adopted for evolutionary research as well. As part of this project, a specification for constructing and computing with phyloreferences within an ontological framework will be researched, implemented, and tested. Phyloreferences will be designed with the goal that any element, whether node, branch, or clades, on the Tree of Life can be referenced in a way that is unambiguous and has fully computable semantics defined by patterns of evolutionary relatedness. The main objectives of this research include creating a formal specification for phyloreference and phylogeny encoding and reasoning in OWL; ascertaining correctness of the specification using small-scale tests verifiable by domain experts; and finally scaling the approach to a large-scale biodiversity data resource navigation proof-of-concept application. Results of the project will be available at http://www.phyloref.org.