

Phyloinformatics in the age of Wikipedia

Roderic D M Page
DEEB, FBLS
University of Glasgow, Glasgow G12 8QQ, UK
r.page@bio.gla.ac.uk

One of the great challenges of phyloinformatics is linking together information on phylogenies, taxa, genomes, specimens, and publications. One approach to linking disparate data is to use shared identifiers. For example, if a bibliographic database and a nomenclatorial database both use the same identifier for a publication (such as a DOI), then we can easily link the two pieces of information together using that identifier. An obstacle to this approach is the lack of identifiers, or failure to reuse existing identifiers. Sequences in GenBank may lack bibliographic identifiers, even if the paper in which the sequences were published has an identifier. Most museum specimens lack resolvable identifiers, thwarting linking sequences to their vouchers. This low “link density” is a major obstacle to linked data approaches to integrating biodiversity data.

This talk will explore two approaches to addressing these problems. The first, <http://bioguid.info>, provides numerous services, including tools for discovering identifiers for publications based on their bibliographic details, and for extracting and resolving specimen identifiers from GenBank sequence records. The second approach uses a semantically-enabled wiki (e.g., <http://iphylo.org/treebase>) to provide tools to annotate information on taxa, specimens, sequences, publications, and phylogenies. These annotations can increase link density (for example, by correcting or updating existing metadata), enabling more sophisticated ways to query and navigate phyloinformatic data.

The talk will also explore wider implications of wiki-style approaches in biodiversity informatics, notably the increasingly important role of Wikipedia (and its derivative DBpedia), and the possibilities of using wikis to crowd-source cleaning up legacy literature being scanned by the Biodiversity Heritage Library.