BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 2: e25776
doi: 10.3897/biss.2.25776

OPEN ACCESS

Conference Abstract

All the Clades in the World: Building a Semantically-Rich and Testable Ontology of Phylogenetic Clade Definitions

Gaurav Vaidya[‡], Guanyang Zhang[‡], Hilmar Lapp[§], Nico Cellinese[‡]

[‡] University of Florida, Gainesville, Florida, United States of America
[§] Duke University, Durham, North Carolina, United States of America

Corresponding author: Gaurav Vaidya (gaurav@ggvaidya.com)

Received: 12 Apr 2018 | Published: 21 May 2018

Citation: Vaidya G, Zhang G, Lapp H, Cellinese N (2018) All the Clades in the World: Building a Semantically-Rich and Testable Ontology of Phylogenetic Clade Definitions. Biodiversity Information Science and Standards 2: e25776. https://doi.org/10.3897/biss.2.25776

Abstract

Taxonomic names are ambiguous as identifiers of biodiversity data, as they refer to a particular concept of a taxon in an expert's mind (Kennedy et al. 2005). This ambiguity is particularly problematic when attempting to reconcile taxonomic names from disparate sources with clades on a phylogeny. Currently, such reconciliation requires expert interpretation, which is necessarily subjective, difficult to reproduce, and refractory to scaling. In contrast, phylogenetic clade definitions are a well-developed method for unambiguously defining the semantics of a clade concept in terms of shared evolutionary ancestry (Queiroz and Gauthier 1990, Queiroz and Gauthier 1994), and these semantics allow locating clades on any phylogeny. Although a few software tools have been created for resolving clade definitions, including for definitions expressed in the Mathematical Markup Language (e.g. *Names on Nodes* in Keesey 2007) and as lists of GenBank accession numbers (e.g. *mor* in Hibbett et al. 2005), these are application-specific representations that do not provide formal definitions with well-defined semantics for every component of a clade definition. Being able to create such machine-interpretable definitions would allow computers to store, compare, distribute and resolve semantically-rich clade definitions.

To this end, the Phyloreferencing project (http://phyloref.org, Cellinese and Lapp 2015) is working on a specification for encoding phylogenetic clade definitions as ontologies using the Web Ontology Language (OWL in W3C OWL Working Group 2012). Our specification allows the semantics of these definitions, which we call phyloreferences, to be described in terms of shared ancestor and excluded lineage properties. The aim of this effort is to allow any OWL-DL reasoner to resolve phyloreferences on a phylogeny that has itself been translated into a compatible OWL representation. We have developed a workflow that allows us to curate phyloreferences from phylogenetic clade definitions published in natural language, and to resolve the curated phyloreference against the phylogeny upon which the definition was originally created, allowing us to validate that the phyloreference reflects the authors' original intent. We have started work on curating dozens of phyloreferences from publications and the clade definition database RegNum (http://phyloregnum.org), which will provide an online catalog of all clade definitions that are part of the Phylonym Volume, to be published together with the PhyloCode (https://www.ohio.edu/phylocode/). We will comprehensively curate these definitions into a reusable and fully computable ontology of phyloreferences.

In our presentation, we will provide an overview of phyloreferencing and will describe the model and workflow we use to encode clade definitions in OWL, based on concepts and terms taken from the Comparative Data Analysis Ontology (Prosdocimi et al. 2009), Darwin-SW (Baskauf and Webb 2016) and Darwin Core (Wieczorek et al. 2012). We will demonstrate how phyloreferences can be visualized, resolved and tested on the phylogeny that they were originally described on, and how they resolve on one of the largest synthetic phylogenies available, the Open Tree of Life (Hinchliff et al. 2015). We will conclude with a discussion of the problems we faced in referring to taxonomic units in phylogenies, which is one of the key challenges in enabling better integration of phylogenetic information into biodiversity analyses.

Keywords

phylogenetics, clade definitions, ontologies, ontology development, phyloreferences

Presenting author

Gaurav Vaidya

Grant title

The Phyloreferencing project is funded by the US National Science Foundation through collaborative grants DBI-1458484 and DBI-1458604 to Hilmar Lapp (Duke University) and Nico Cellinese (University of Florida), respectively. The proposal text is available online (Cellinese and Lapp 2015).

References

- Baskauf S, Webb C (2016) Darwin-SW: Darwin Core-based terms for expressing biodiversity data as RDF. *Semantic Web* 7 (6): 629-643. https://doi.org/10.3233/sw-150203
- Cellinese N, Lapp H (2015) An ontology-based system for querying life in a post-taxonomic age. Figshare https://doi.org/10.6084/M9.FIGSHARE.1401984
- Hibbett D, Nilsson RH, Snyder M, Fonseca M, Costanzo J, Shonfeld M (2005) Automated phylogenetic taxonomy: An example in the Homobasidiomycetes (mushroom-forming fungi). *Systematic Biology* 54 (4): 660-668. https://doi.org/10.1080/10635150590947104
- Hinchliff C, Smith S, Allman J, Burleigh JG, Chaudhary R, Coghill L, Crandall K, Deng J, Drew B, Gazis R, Gude K, Hibbett D, Katz L, Laughinghouse HD, McTavish EJ, Midford P, Owen C, Ree R, Rees J, Soltis D, Williams T, Cranston K (2015) Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proceedings of the National Academy of Sciences* 112 (41): 12764-12769. https://doi.org/10.1073/pnas.1423041112
- Keesey TM (2007) A mathematical approach to defining clade names, with potential applications to computer storage and processing. *Zoologica Scripta* 36 (6): 607-621. https://doi.org/10.1111/j.1463-6409.2007.00302.x
- Kennedy J, Kukla R, Paterson T (2005) Scientific names are ambiguous as identifiers for biological taxa: their context and definition are required for accurate data integration. In: Ludäscher B., Raschid L. (eds) Data Integration in the Life Sciences. DILS 2005. *Lecture Notes in Computer Science* 3615: 80-95. https://doi.org/10.1007/11530084_8
- Prosdocimi F, Chisham B, Pontelli E, Thompson JD, Stoltzfus A (2009) Initial implementation of a Comparative Data Analysis Ontology. *Evolutionary Bioinformatics Online* 5: 47-66. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2747124/
- Queiroz Kd, Gauthier J (1990) Phylogeny as a central principle in taxonomy: phylogenetic definitions of taxon names. *Systematic Zoology* 39 (4): 307. https://doi.org/10.2307/2992353
- Queiroz Kd, Gauthier J (1994) Toward a phylogenetic system of biological nomenclature. *Trends in Ecology & Evolution* 9 (1): 27-31. https://doi.org/10.1016/0169-5347(94)90231-3
- W3C OWL Working Group (2012) OWL 2 Web Ontology Language, document overview (second edition). https://www.w3.org/TR/2012/REC-owl2-overview-20121211/. Accessed on: 2018-3-26.
- Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, Robertson T, Vieglais D (2012) Darwin Core: an evolving community-developed biodiversity data standard. *PLOS ONE* 7 (1): e29715. https://doi.org/10.1371/journal.pone.0029715