Building an ontology of logic definitions for groups of biological organisms to enable data integration

Gaurav Vaidya[1], Hilmar Lapp[2], Nico Cellinese[1]

[1] University of Florida and Florida Museum of Natural History

[2] Duke University

All organisms are related to each other: an example with alligators

Biologists visualize hypotheses of how organisms are related to each other using *phylogenetic trees*. For example, the following phylogenetic tree includes the hypothesis that alligators (Alligatoroidea) and crocodiles (Crocodylidae) are more closely related to each other than either are to gavials (*Gavialis gangeticus*).

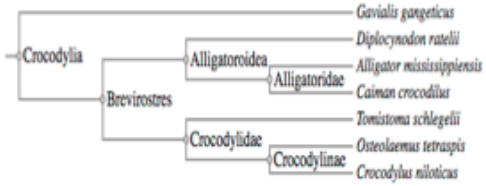
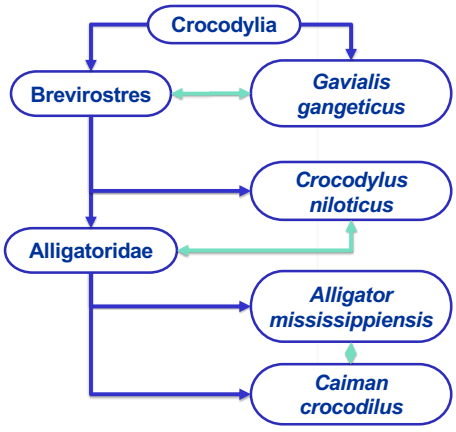


Fig 1. Phylogenetic tree of alligators, crocodiles and caimans from Brochu 2003 [1].

Phylogenetic trees can be represented in OWL using CDAO

The Comparative Data Analysis Ontology (CDAO) [2] allows phylogenetic trees to be represented as an ontology in the Web Ontology Language (OWL).



has Child	Transitive as *has Descendant*
has Sibling	Symmetrical

Note that nodes on this phylogenetic tree represent biological groups (*taxa*). We can relate these nodes to taxonomic names using the *represents TU* property from CDAO [2].

Node *Alligator mississippiensis*	*represents TU* **some** (*Has Name* **some** (*Zoological name* **and** *Has Scientific Name* **value** "Alligator mississippiensis"))
Node *Crocodylus niloticus*	*represents TU* **some** (*Has Name* **some** (*Zoological name* **and** *Has Scientific Name* **value** "Crocodylus niloticus"))

Phylogenetic clade definitions identify groups of organisms unambiguously

Phylogenetic clade definitions [3] define groups in terms of their ancestral relationships by specifying biological groups that must be either included in or excluded from the clade on a particular phylogenetic tree:

Alligatorinae	*Alligator mississippiensis* and all crocodylians closer to it than to *Caiman crocodilus*.
Brevirostres	Last common ancestor of *Alligator mississippiensis* and *Crocodylus niloticus* and all of its descendents.

How can we translate a definition based on ancestral relationships into an OWL restriction expression?

We can use **OWL Property Chains** as used in [4] to create clade definitions in OWL to traverse phylogenetic trees:

- *includes TU*: *has Descendant* **o** *represents TU*
- *excludes TU*: *has Sibling* **o** *has Descendant* **o** *represents TU*

We can build an OWL restriction that represents a clade definition with one inclusion and one exclusion using these two property chains:

Name	OWL restriction	Nodes matched
Alligatorinae	*includes TU* **some** *Alligator mississippiensis*	*Alligator mississippiensis*, Alligatoridae, Brevirostres, Crocodylia
	and *excludes TU* **some** *Caiman crocodilus*	*Alligator mississippiensis*, *Crocodylus niloticus*, *Gavialis gangeticus*

The parent of this node will necessarily be the smallest clade that includes both specifiers. We can use this to construct an OWL restriction for a clade definition that includes two taxa:

Name	OWL restriction	Nodes matched
Brevirostres	*has Child* **some**	Brevirostres
	(*includes TU* **some** *Alligator mississippiensis*	*Alligator mississippiensis*, Alligatoridae, Brevirostres, Crocodylia
	and *excludes TU* **some** *Crocodylus niloticus*)	Alligatoridae, *Gavialis gangeticus*

More complex clade definitions require chaining OWL restrictions together and testing different ways in which clades could be related to each other.

Producing an ontology of clade definitions

We curate clade definitions into a JSON-LD representation of an OWL ontology.

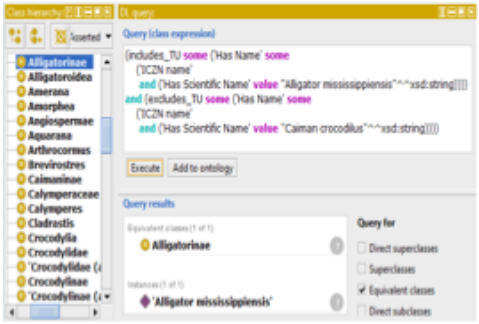


Fig 2. A clade definition in Protégé [5].

Finding a clade on the Open Tree of Life

The Open Tree of Life [6] synthesizes information from 987 phylogenies and several taxonomic resources to provide a draft hypothesis of evolutionary relationships between 2,640,941 taxa. Resolving clade definitions on the Open Tree of Life would facilitate navigation.

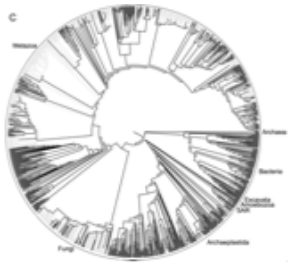


Fig 3. Lineages on the Open Tree of Life synthetic phylogenetic tree as visualized in [6].

Acknowledgments

The Phyloreferencing project is funded by the US NSF through collaborative grants DBI-1458484 (HL) and DBI-1458604 (NC). GV's attendance of US2TS was funded by the US2TS Travel Grant.

References

1.Brochu (2003) Ann Rev Earth Plan Sci 2003 31:1, 357-397. https://doi.org/10.1146/annurev.earth.31.100901.141308

2.Prosdocimi *et al.* (2009) Evol Bioinform Online. 5:47–66. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2747124/

3.de Queiroz and Gauthier (1990) Syst Biol 39(4):307–322. https://doi.org/10.2307/2992353

4.Carral *et al.* (2017) Arxiv 1710.05096 https://arxiv.org/abs/1710.05096

5.Musen, M.A.ACM SIG AI 1(4) http://dx.doi.org/10.1145/2557001.25757003

6.Hinchliff *et al.* (2015) Proc Nat Acad Sci 112.41 (2015): 12764-12769. https://doi.org/10.1073/pnas.1423041112