Integrating Anatomy and Phenotype Ontologies with Taxonomic Hierarchies

James P. Balhoff^{1,2}, Peter E. Midford¹, Hilmar Lapp¹

¹National Evolutionary Synthesis Center, Durham, NC, USA ²University of North Carolina, Chapel Hill, NC, USA

Cross-species anatomy ontologies and shared organismal taxonomies provide powerful tools for integrating comparative observational data with other knowledge domains. In current practice, organismal taxonomies are typically represented as subclass hierarchies in which taxa are modeled as ontological classes and individual organisms as instances of those classes. However, in contrast to the typological nature of ontological classes, evolutionary data regularly show polymorphism, reversals, and novelty. We have found that modeling taxonomic groups as individuals instead of as classes is more consistent with evolutionary biological data and knowledge, and also has desirable practical consequences when linking observational data such as phenotypes to taxa within the OWL DL framework.

The way taxonomies group organisms into successively lower and lower hierarchical levels is informed by the characteristics that the members of these groups share, and those that distinguish them from each other. It is thus tempting to view a taxonomic hierarchy as a genus-differentia system in which classes can be logically distinguished from one another by the essential features that their members (i.e., instances) must possess. For example, "Tetrapoda are vertebrates that have four limbs", or, "Mammalia are tetrapods that have hair and mammary glands". However, the features shared by organisms grouped together in this way are the result of their presumed common evolutionary descent, and it is foremost this shared evolutionary history that organismal taxonomies aim to reflect. Hence, the distinguishing features of a group are assumed to have originated in their common ancestor, but may subsequently among the lineage of descendants become modified, lost, or polymorphic due to the continuing evolutionary process. As David Hull writes, "Organisms belong in a particular species because they are

part of that genealogical nexus, not because they possess any essential traits" [1].

As a consequence, applying a naive typological approach to modeling evolutionary data in a Description Logic (DL) framework such as OWL DL can result in unintended inferences and logical inconsistencies. For example, it is cumbersome to represent polymorphic character states, and expressing evolutionary reversals or loss of phenotypes using cardinality or negation ('has_part only (not scale)') can result in unsatisfiable classes. If instead we model taxa as individuals, and use object properties rather than subsumption to reflect the taxonomic hierarchy, we can control the propagation of property assertions up and down the taxonomic hierarchy.

For the purposes of demonstrating the advantages of this approach in several examples, we use the transitive object properties subclade_of (and its inverse contains_clade) to establish the hierarchy of taxa, which are all instances of type Clade, and the object property member_of (and its inverse has_member) to relate individual organisms to a taxon. We focus our examples on anatomical phenotypic traits, which we express as (anatomical) entities that bear some quality, and associate to organisms using the has part relationship. Using these conventions, we can capture the observation of a bifurcated dorsal fin in the channel catfish Ictalurus punctatus in the following way:

```
Individual: Ictalurus_punctatus
Types: Clade, has_member some
  (has_part some (caudal_fin and
  (bearer_of some bifurcated)))
Facts: has_rank species,
  subclade of Ictalurus
```

By defining a property chain $contains_clade$ o $has_member \rightarrow has_member$, we can infer that the order Siluriformes, which $contains_clade$ Ictalurus (and thus Ictalurus punctatus), has

some member organism with a bifurcated caudal fin. Additionally the set of all such orders can easily be expressed as '(has_rank value order) and (has_member some (has_part some (caudal_fin and (bearer_of some bifurcated))))'. Thus, querying for higher taxa based on traits identified in sub-taxa, or in other words up-propagating trait observations to higher taxa, is straightforward in this approach, but difficult to accomplish if taxa are modeled as classes.

In addition to direct observations of traits, a common result of comparative analyses is the assertion of certain shared traits as ancestral to a group or organisms with shared evolutionary history. Therefore, to be consistent with biological knowledge such phenotypes should be asserted only once for the ancestor, and then be inferred (down-propagated) for each descendant organism, and the taxa that these are members of. We can accomplish this by introducing the object properties has ancestor (and its inverse *has_descendant*), which relates organisms to their ancestors, and has progenitor, a subproperty of has_member that relates a taxon to the ancestral member of that taxon. For example, Siluriformes have lost scales, which we can express in the following way:

```
Individual: Siluriformes
Types: Clade, has_progenitor some
  (has_part only (not scale))
Facts: has rank order
```

By defining the property chain 'subclade_of o has_progenitor \rightarrow has_ancestor_to_members', we can straightforwardly express the set of all species with an ancestor that does not have scales as:

```
((has_rank value species) and (has_ancestor_to_members' some (has part only (not scale)))).
```

Traits are borne by, and observed for individual organisms. To facilitate free inter-

mingling of and reasoning across directly observed traits with those hypothesized as ancestral states, they should be expressed consistently as properties of organisms, rather than those of taxa. The individual-based approach we propose here accomplishes this in a straightforward manner.

Some of the presented issues can also be addressed within a class-based taxon model, for example by adding a level of indirection between the taxon class and the phenotype assertion. However, such approaches still make it difficult to appropriately propagate phenotypic data across levels of the taxonomy as shown above. Furthermore, assigning population-specific property values, such as geographic range or temporal extent, in a class-based approach suffers from similar problems, whereas it is straightforward in an individual-based modeling approach.

conclusion, In representing taxa ontological classes presents an unrealistic model of organisms as instances of kinds. which creates a variety of problems for ontologically modeling observations of organismal properties. Embracing instead "population thinking" [2], and correspondingly modeling taxa, organisms, and their observed traits in a way that directly represents the historical individuality of evolutionary lineages could integration and reasoning comparative observational data in a fully OWL DL compatible framework not only much more straightforward, but also more consistent with our understanding of biological evolution.

References

- Hull, D.L. A matter of individuality. Philosophy of Science 45(3), 335—360 (1978)
- Mayr, E.. Populations, species, and evolution. Harvard University Press, Cambridge, MA (1970)