Project Summary

Overview: The rich and continuously growing treasure trove of morphological data published in the literature is key for understanding the diversity of life on earth. Although a rich ecosystem of computational tools exists for managing and analyzing morphological traits comparatively, the ability to compute with morphological data in direct connection with the accumulated body of morphological knowledge has remained elusive. There is no readily accessible computational infrastructure that tools for comparative trait analysis can tap into to understand what character and character state descriptions mean given the requisite domain knowledge, and thus to objectively and reproducibly assess the relatedness, independence, and distinctness of characters and character states. This is in spite of the enormous breakthroughs that have been made over recent years in powerful and efficient technologies for knowledge representation, discovery, and machine reasoning. However, deploying and leveraging these capabilities requires intensive upfront investments, such as building infrastructure and training a workforce, that are out of reach for most investigators. The resulting opportunity cost is high. While ontologies have been used with great success in knowledge domains across the life sciences, the considerable investment in the development of rich ontologies, including for morphology and organismal traits across the tree of life, remains largely unexploited in evolutionary science.

This proposal aims to address this gap by creating a centralized computational infrastructure that affords comparative analysis tools the ability to compute with morphological knowledge through scalable online application programming interfaces (APIs). To accomplish this, the project will adapt key products and know-how developed by the Phenoscape project, including an integrative knowledgebase of ontology-linked phenotype data, metrics for quantifying the semantic similarity of phenotype descriptions, and a nascent API.

Intellectual Merit: This work adapts a large body of technology development work and engineering know-how to transform the computational capabilities available to users and developers of tools for comparative trait analysis. The concrete objectives focus on addressing three long-standing needs for which the difficulty of computing with domain knowledge is the major impediment: (1) computationally synthesizing, calibrating, and assessing morphological trait matrices from across studies; (2) objectively and reproducibly incorporating morphological domain knowledge provided by ontologies into evolutionary models of trait evolution; and (3) generating testable hypotheses for adaptive diversification by incorporating semantic phenotypes into ancestral state reconstruction and identifying domain ontology concepts linked to evolutionary changes in a branch or clade more frequently than expected by chance.

Broader Impacts: Similarly to how easy access to sophisticated machine learning and artificial intelligence capabilities over online APIs has transformed the capabilities of mobile device applications, the computable domain knowledge capacities enabled by this project will transform how tools can derive insight from morphological big data. To prepare potential users and developers of comparative analysis tools for adopting these capabilities, participants in this project will develop the curriculum for and teach a short-course on requisite knowledge representation and computational inference technologies, tailored to evolutionary biologists. The effectiveness of the training will be assessed and successively improved. To promote adoption of, and innovative applications using the capabilities created by the project, hackathon events will be held to bring together comparative tool developers with users whose research questions could benefit from these new capabilities.