Project Description

OVERVIEW

Most biological data and knowledge are directly or indirectly linked to biological taxa. Linking data by the taxon they have in common is one of the most fundamental and ubiquitous ways a wide range of biological data are integrated, aggregated, and indexed, from genomic and microbial diversity to macro-ecological data. Yet, to this day most methods and resources developed for this purpose operate on the basis of Linnaean nomenclature. Although taxon names often work well enough for well established, stable, and homogenous taxa (such as *Homo sapiens*, *Mus musculus*, and other model organisms), they lead to numerous problems when applied to biodiversity at large. Chiefly, because they are text strings taxon names are impoverished in machine-accessible semantics that could be used for computation. Imagine, for comparison, that for geolocation-linked data only place names, not standard latitude/ longitude geo-coordinates, were available for computation. Data could not be aggregated by region, users could not draw a bounding box on a map to query a database, species occurrence data could not be queried for "all species within 50 miles of my location", and users querying by place would have to know country, state, and possibly city to make the query less ambiguous. Yet, this is the situation in computing with taxon-linked data. As a result, computation-reliant science involving biodiversity data is hampered by a number of challenges. These include (but are not limited to) the ambiguity of Linnaean taxon names due to the concept as which they are applied being neither fixed nor accessible to machines; the difficulty of referring to nameless clades in a phylogeny in any kind of consistent or reproducible way; and referring to clades in a persistent way when phylogenetic knowledge is updated. These challenges will remain even in light of grand synthesis efforts for a unified and comprehensive Tree of Life coming to fruition.

We aim to address these by developing, testing, and prototyping an approach, called *phyloreferencing*, to defining any group of interest in a phylogeny such that the semantics of the definition is fully accessible to machines. In this approach, a *phyloreference* is a formal expression in the Web Ontology Language (OWL) [1], the semantics of which are well defined [2] and available to computation via OWL reasoners. Ontology and reasoning technologies have already shown their power for biological knowledge integration and discovery [3-5] and are increasingly being adopted for evolutionary research as well [6-9]. Although in principle any type of ontologically expressible property could be used for computable semantics, we focus specifically on the one property common to all of life, the pattern of evolutionary descent. As such, our approach builds on a considerable body of prior work on phylogenetic taxon definitions, both theoretical [10,11] as well as applied [12,13] (also see Background), and follows on earlier but so far disparate ideas and initiatives [14-17], including one co-lead by the PIs [17]. Given a phyloreference and a phylogeny represented in OWL, any general purpose OWL reasoner can infer which elements in the phylogeny match (formally, subclass or instantiate) the phyloreference, which in analogy to geographic information systems, we call *resolving a phyloreference*. A phyloreference compliant with our approach bears the following key properties. (1) Because it consists of uniquely identified ontology terms and properties, it is unambiguous. (2) Although it expresses a pattern of shared ancestry, it can be defined and communicated independent of a concrete phylogeny. (3) It may be named (labeled), but a name only aids communication and carries no semantics. (4) Its semantics are interpretable by any off-the-shelf OWL reasoner implementation, and do not require custom, bespoke tools. (5) To promote reuse and consistency of frequently used phyloreferences, they can be compiled and published in the form of community-vetted OWL ontologies. (6) Tools and algorithms exist that use the ontologies from which the terms used within a phyloreference are drawn to compute quantitative metrics between phyloreferences, including distance and semantic similarity [8,18-20].

Primary use-cases. To illustrate some of the current challenges with taxon names for organism-linked data integration, consider the taxon Campanulaceae, a large flowering plant clade in the Asterales (and one of the study subjects of PI Cellinese). The concept attached to the name changed multiple times since Jussieu first established the family in 1789. Its application went from much broader (e.g., [21,22]), when a group that now forms a separate family (Pentaphragmataceae) was included, to much narrower (e.g., [23-25]), when it encompassed only the genera now more commonly referred to as the Campanuloideae [26]. Nonetheless, even in the most recent literature some authors continue to apply the name Campanulaceae in the narrow sense instead of using the name Campanuloideae (e.g., [27-29]), while others apply it to include four lineages (Lobelioideae, Cyphioideae, Nemacladoideae, Cyphocarpoideae) in addition to Campanuloideae (e.g., [30,31]). To make matters worse, the name Campanuloideae has also been applied to a subclade within the lineage that is normally designated by this name (e.g., [32], Figure 1b). The Global Biodiversity Informatics Facility (GBIF) [33], one of the largest aggregator of species occurrence data, does not currently have Campanuloideae in its backbone taxonomy (though it does in two external taxonomies that it imports), and as a result it has no data linked to it. Both the Encyclopedia of Life (EOL) [34], another global biodiversity data aggregator, and the Open Tree of Life [35], a phylogeny synthesized from published phylogenetic knowledge, suggest that Campanuloideae is a misspelling of Campaniloidea, a completely unrelated group.

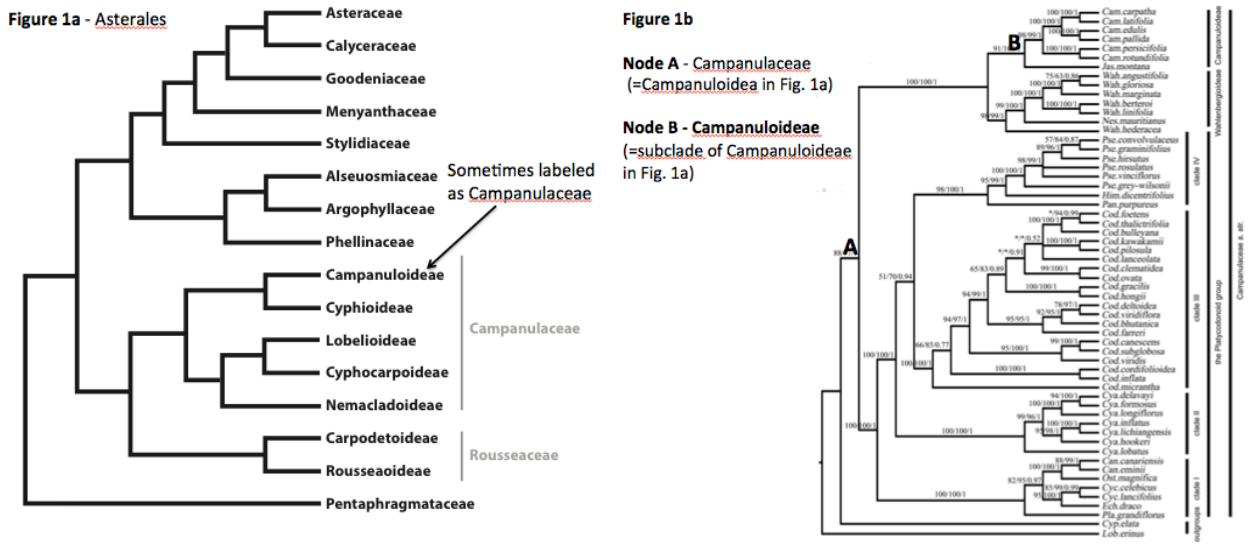


Figure 1: Phylogeny of Asterales and Campanulacaea

A phyloreference for Campanuloidea would look similar to this (OWL Manchester Syntax; properties in italics; for readability, the ontologies of constituent terms and properties are omitted, and term labels are used in place of identifiers): `<Campanuloidea> EquivalentTo` *`has_Descendant`* `value Campanula_latifolia and` *`excludes_lineage_to`* `value Lobelia`. This expression, which models a branch-based phylogenetic definition (see Background, Figure 2) and is a valid axiom in OWL, asserts that the class `Campanuloidea` is equivalent to the most inclusive node from which 'Campanula_latifolia' is descended, and which excludes the lineage leading to node 'Lobelia', two necessary and sufficient conditions. The definition of the property *`excludes_lineage_to`* ensures that it is only the most inclusive node that matches. The semantics of such a definition are transparent, unambiguous, and readable by any OWL-aware tool, including machine reasoners. The definition does not pinpoint one particular node in one particular taxonomy or phylogeny. Instead, it can be applied to, i.e., a reasoner can infer the matching node(s) in, any taxonomy or phylogeny (in the form of an

ontology) so long as the two necessary and sufficient conditions are inferable. For example, if a target phylogeny lacks node `Campanula_latifolia` but contains `Campanula`, a (mapping) axiom asserting `Campanula` *`has_Descendant`* `Campanula_latifolia` maintains a match of the condition inferable (due to transitivity of *`has_Descendant`*). In practice, creating effective mapping axioms is non-trivial and part of our proposed work is developing solutions for it.

Another challenge for organism-linked data is presented by polyphyletic species, which are very common across the domains of life due to hybridization, incomplete lineage sorting, or ancient genome duplication events. For example, PI Cellinese's lab has recently found that *Campanula erinus*, a widespread taxon in the Mediterranean basin nested in a clade of narrow Aegean islands endemics, is highly polyphyletic and a polyploid. However, clades can be recovered with interesting morphological or genetic synapomorphies (unpubl. results). Until a formal revision of the taxon, which is a slow process, taxon names offer no help in referring to such clades, especially if, as it is very common, the type specimen is missing from the analysis. In other domains, in particular bacteria and viruses, taxa are often so poorly known that only unnamed phylotypes can be identified (e.g., [36]). These phylotpes are often referred to as 'dark taxa' [37,38]. Phyloreferences can address these cases, because any uniquely identifiable object can be used to identify the nodes to which the conditions in a phyloreference refer. To illustrate this point, the above example is simplified in assuming that a node directly identified as `Campanula_latifolia` exists. However, a node could equally well be identified as those matching certain properties, for example *`has_Descendant`* `some (has_accession_number value "Genbank: EF141027").`

Major Objectives and Deliverables. Our development plan is designed to accomplish 3 main objectives: creating a formal specification for phyloreference encoding and reasoning in OWL; ascertaining correctness of the specification using several use cases verifiable by domain experts; and finally scaling the approach to a large-scale biodiversity data resource navigation proof-of-concept application. Specifically, our plan consists of the following deliverables.

1. **A specification for encoding phyloreferences and phylogenies in OWL.** We will specify and test templates and a supporting ontology for constructing phyloreferences in OWL, guided by phylogenetic definitions used in the literature. In parallel, we will develop a model and an automatic transformation tool for representing phylogenies in OWL such that phyloreferences can be resolved by OWL reasoning.
2. **An OWL ontology of vetted phyloreferences.** To ground-proof the specification, we will create a tool to transform the published phylogenetic definitions contained in the RegNum database [39] to an OWL ontology of phyloreferences. We will also supplement the RegNum content with phylogenetic definitions culled from the angiosperm (flowering plants) systematics literature.
3. **A proof-of-concept application for utility and correctness of phyloreferences.** Using a comprehensive phylogenetic tree for angiosperms and the previously curated phyloreferences, we will create an online application that uses OWL reasoning to allow users to query the tree using phyloreferences, and to find phyloreferences based on chosen nodes of the tree.
4. **A proof-of-concept application for navigating large-scale data resources.** We will extend the proof-of-concept application to allow users to query and navigate EOL with phyloreferences, using the full synthetic Open Tree of Life.

Significance. The outcomes of this project stand to benefit the entire biological community, by laying the foundation for an informatics infrastructure that enables using the Tree of Life to organize, query, and navigate our knowledge of the diversity of life. Our work is particularly timely in this regard. Until only recently, there were tens of thousands small, disparate, incongruent, and difficult to align phylogenies, which still covered only a small minority of life.

However, this is changing rapidly. Within the last decade, an increasing number of large phylogenies with up to tens of thousands of tips encompassing large and diverse groups have been published (e.g., [40,41]), and since 2012 the NSF-funded Open Tree of Life project [35] endeavors to produce a single synthetic phylogeny for all of life. For such large tree synthesis projects the need for phyloreferencing has already arisen [16,35], though specific to attaching annotations to elements of the tree, for example node ages. Although our approach could be used for such purposes, too, our ultimate goal is enabling machine-based integration and querying of practically any organism-linked data by patterns of shared descent.

There are also parts of the Tree of Life for which a stunning organismal biodiversity is only just beginning to be characterized, and for which the traditional fallback of Linnaean names is hardly available, and perhaps never will be. For such parts, for example the microbial domain [42,43], newly discovered taxa are known only as phylotypes or other discretionary names. Yet, the ability to unambiguously refer to these groups is necessary, not least to organize, query, and retrieve our knowledge about any group of interest. In contrast to Linnaean names, phylogenetic definitions can be created using any identifiable object, including specimens, samples, and sequences. If appropriately labeled and distributed in community-vetted ontologies, phyloreferences can provide names that allow researchers to communicate data and knowledge about their groups, yet also have fully computable and thus reproducible semantics built in. This potential of phyloreferences extends below the species level, for example to label and query monophyletic entities corresponding to subsets of populations or polyploid derivatives that show interesting evolutionary and/or biogeographic patterns but are currently unnamed and therefore, cannot be discovered. These entities are not considered ‘species’ and a clear mechanism to name them is lacking from all of the formal nomenclature codes.

Our work is timely also because as a result of new phylogenetic studies having become more affordable, the amount of phylogenetic taxon definitions being published has increased rapidly in recent years across multiple domains ([44-57] and numerous others). This signifies that phylogenetic approaches to defining taxonomic groups and their names are being increasingly widely adopted. Our work on phyloreferences can eventually result in a community standard that allows these definitions, including their semantics, to be fully accessible to machines, rather than remaining buried in the text of publications.

BACKGROUND

Prior to Linnaeus, biological knowledge was organized in large, poorly defined categories, and nomenclature was completely unstructured. Linnaeus brought order by generating criteria to define logical relationships among abstract classes (categorical ranks) and restructuring the nomenclatural system by assigning a binomen to every organism at the species level and a single name to every higher rank. Outside of the yet to be established unifying context of evolution, taxa were assumed to be static entities, with character similarity providing the best approach to defining groups of organisms. Linnaean nomenclature served the need of linking names to these taxon groups. Darwinian theory revolutionized the perspective on biological relationships and taxon group membership to the notion that it is natural processes that give rise to taxa, whilst characters can only diagnose, and not define categories. Hennig formalized Darwin’s theories and provided the criteria to construct phylogenetic trees [58]. Although both Hennig and Linnaeus proposed hierarchical frameworks to represent biological knowledge, the phylogeny-governed hierarchy in Hennig’s framework reflects genealogical relationships, as opposed to the logical relatedness of groups based on arbitrary ranks in Linnaeus’ framework. While tree-thinking slowly displaced the phenetics approach, phylogenetic systematics until the 1990's remained largely decoupled from biological nomenclature, in part because of the need to maintain classifications stable [59]. As a consequence, Linnaean nomenclature has subsequently been repurposed to link names to clades, nodes and terminal lineages [60].

However, names that point to traditionally defined taxon groups may only approximately correspond to clades [60].

Starting in the mid 1980's a number of authors suggested that taxon names could be defined by reference to a part of a phylogenetic tree, prompting an extensive theoretical discussion as well as the first attempts to generate phylogenetic definitions [12,13,59,61-65]. This body of work laid the foundation for phylogenetic taxonomy, later renamed to phylogenetic nomenclature, which takes a strictly tree-thinking approach to biological nomenclature [10,11,66]. Soon thereafter, the *PhyloCode* [67] was drafted as an application of phylogenetic nomenclature principles. Many early and subsequent systematics papers [68-103] that clearly articulated the need to communicate parts of the Tree of Life applied phylogenetic nomenclature to selected taxon groups, leading to the emergence of three basic clade types and their associated phylogenetic definitions. These are (1) node-based definitions, denoting a clade that includes the most recent common ancestor, and all its descendants, of two or more specified ingroup nodes; (2) branch (or stem)-based definitions, denoting a clade that includes the branch (or stem) subtending the first ancestor, and all its descendants, of one specified node in the "ingroup" but excluding the ancestor(s) of any specified "outgroup" nodes; and (3) apomorphy-based definitions, denoting the clade that arises from the first ancestor, and includes all its descendants, that possesses a specified character that is synapomorphic with a taxon in the ingroup (Figure 2).

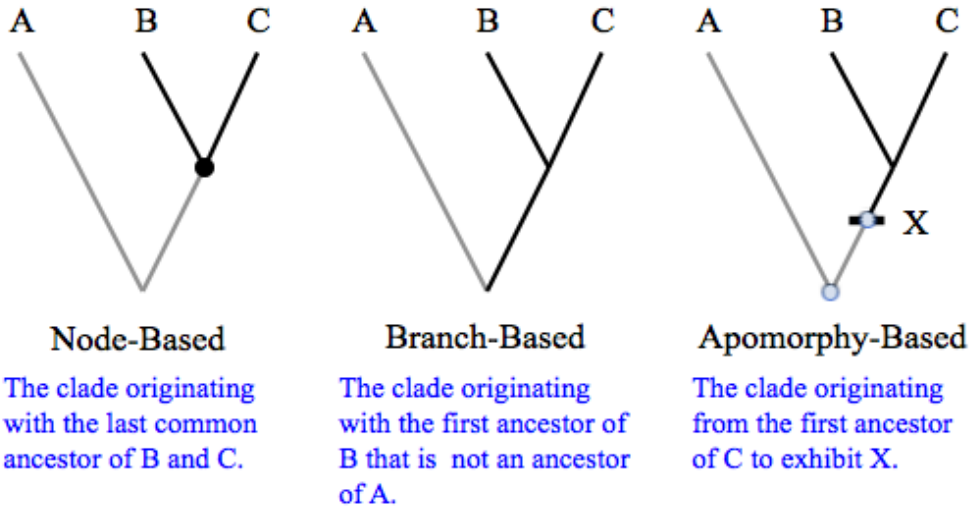


Figure 2. The 3 main types of phylogenetic definitions.

Just as georeferences are not an instrument for naming places, phyloreferencing as proposed here is not meant to promote phylogenetic over Linnaean nomenclature. Instead, we seek to construct an informatics framework for integrating and navigating organism-linked data by concepts not afforded by Linnaean taxonomies, by building on the theoretical as well as applied results from a wealth of earlier work, including on phylogenetic nomenclature. We are not the first to do so. Keesey [14] and in part Sereno [15,104] have already envisioned mechanisms and applications that would leverage computable clade definitions to unambiguously retrieve taxa based on shared descent-based specifications. Keesey even includes a notation and formalism for defining clade names based on mathematical set theory and operators. However, not using a standard ontology framework requires him to define custom semantics for the notations, the formalism is unfamiliar even among technically savvy domain scientists, and easy-to-apply off-the-shelf tools to edit or compute with expressions in his proposed format are scarce. Perhaps primarily for these reasons his proposal has so far not been adopted.

In contrast, our phyloreferencing approach takes place within an OWL2 framework [1], specifically within the Description Logic (DL) subset of OWL2 (OWL2-DL in short) [2]. Phyloreferences are modeled as defined OWL classes, which means that they have a subclass or equivalence axiom to an OWL class expression that gives, in the form of an intersection of sets each defined by an OWL expression, the necessary and sufficient conditions for class membership. In terms of entailments, this means that any instance (an entity that is not a class) that does not fulfill one or more of the conditions cannot be a member (or instance) of the class, and any instance that fulfills all conditions must be a member of the class. A condition can be the requirement to instantiate (being a member of) a specific class, called a class restriction, or a so-called property restriction, which specifies which kind of value (in the case of object properties) or what particular value (in the case of data properties) an instance must have for the given property to make the restriction true. For example, “Node and (has_Child some Node)” describes all instances of the class Node that also have some instance of Node as value of a has_Child property, i.e., the non-terminal nodes in a tree. Class expressions enable rich reasoning, because OWL reasoners can use these both to infer class membership of instances, and to infer equivalences and subsumption hierarchies between class expressions. Consequently they have been employed with great success in several biological knowledge integration and discovery applications [4,105-107]. Class expressions may remain unnamed; these are oftentimes referred to as post-coordinated or post-composed classes (or terms) [7,108]. They can also be named, by asserting a named class as equivalent to the class expression.

RESULTS FROM PRIOR NSF SUPPORT

Hilmar Lapp is co-PI on NSF DBI-1062404, “ABI Development: Ontology-enabled reasoning across phenotypes from evolution and model organisms”, $479,175 to Duke University as subaward from University of North Carolina at Chapel Hill, 7/1/11 – 6/30/15. The project, called Phenoscape [109], is creating an integrated knowledgebase of vertebrate evolutionary and model organism phenotype observations made fully computable by transforming them into ontological expressions in OWL. The ultimate goal of the project is to use OWL reasoning and computational metrics to connect genes from model organisms to evolutionary transitions by the semantic similarity of their phenotypic profiles. Lapp oversees the OWL modeling and reasoning as well as software development. The project has thus far resulted in 10 journal and conference publications; production of new vertebrate-wide ontologies; major contributions to existing community ontologies, including the cross-metazoan anatomy ontology Uberon [107]; a body of ontology-based semantic annotations for evolutionary fin/limb phenotypes curated from the systematics literature; and various software for curation, processing, reasoning, and integration of ontology-based data.. All data and software created by the project are openly developed and available under CC-BY [110], CC0 [111], and MIT licenses, respectively. The project has also successfully run its major broader impact activities, including a Junior Biocurator program (“Project Exploration”) at the University of Chicago (co-PI: Paul Sereno) with hands-on learning activities on comparative anatomy for high-school students from low-income neighborhoods, and undergraduate research experiences in evolutionary developmental biology for American Indian undergraduate students, run at the University of South Dakota (PI: Paula Mabee). Particularly relevant to this proposal, the scale and logical expressivity that the Phenoscape project has to cope with have yielded invaluable insights into the advantages and downsides of different OWL models for phenotype observations [112], as well as performance limitations of different OWL-DL reasoners and RDF triple stores.

Nico Cellinese is a PI on NSF DEB-0953677, "CAREER: Evolution in the Eastern Mediterranean Campanulaceae - Integrating Information Management and Scientific Workflows in Daily Research and Teaching", $865,251 to University of Florida, 3/1/2010-2/28/2015. This

project explores the biological diversity and biogeography of Mediterranean Campanulaceae (bellflowers). Campanulaceae are a large clade of flowering plants that includes five lineages whose relationships are poorly known. As part of this project, large-scale evolutionary patterns have been assessed to better understand processes that have shaped this group in the Mediterranean basin. Additionally, based on phylogenetic data, new clade definitions have been proposed for Campanulaceae and some of its nested lineages [113-115]. Through this work, PI Cellinese has been able to contribute to the understanding and development of the angiosperm tree of life that resulted in a major publication [48], in which the authors propose a number of clade definitions for flowering plants. Overall and so far, this research has contributed to 12 papers that improve our understanding of the systematics and evolution of Campanulaceae, flowering plants, methods in biogeography, and biological nomenclature. The broader impact accomplishments include an exhibit on the evolution of Campanulaceae in the Mediterranean region, which is currently on a 1-year display at the Florida Museum of Natural History, and the development of curricula for two new graduate and undergraduate-level courses on evolutionary biogeography. PI Cellinese has also been involved in a number of other NSF funded projects (DEB 827609, IOS 0827254, DEB 0829313, DEB 0431258, DBI 0956371) that led to development of ToLKIN [116,117] and BiSciCol [118,119] and so far generated four papers on biodiversity informatics and phyloinformatics. All software code generated in the Cellinese Lab is available publicly under open-source licenses at Github [120].

Lapp and Cellinese were among the co-organizers of the 2009 Phyloinformatics VoCamp [121], and were among the co-leaders of the Phyloreferencing subgroup that formed there [17]. A VoCamp is a hands-on codefest-like workshop focused around shared vocabularies and ontologies rather than software tools. The workshop was funded by the National Evolutionary Synthesis Center (NESCent) and thus indirectly by NSF. The Phyloreferencing subgroup examined and documented phyloreference-driven use-cases, queries, and gaps in supporting infrastructure and shared vocabularies. It is also to our knowledge the first documented use of the term.

RESEARCH AND DEVELOPMENT PLAN

We posit that a phyloreference specification should meet the following requirements.

1. Any element of a phylogeny is referenceable. Phyloreferences may be primarily used for nodes, but the formalism should be just as well capable of denoting branches, entire clades, or only the tip nodes included by a clade, etc.
2. Phyloreferences are unambiguous. Given a phyloreference and a phylogeny, the elements of the phylogeny matching the phyloreference are axiomatically determined by the definition of the phyloreference. A mapping may be needed to allow resolution, but must be axiomatic as well, and thus transparent and machine-processable.
3. Phyloreferences are fully computable. A phyloreference can be resolved by a machine. The semantics of a phyloreference, and thus also its changes from one revision to another, are fully accessible to machines for computational processing.
4. Phyloreferences are portable. The definition of a phyloreferences is neutral with respect to the phylogeny against which one chooses to apply it, and is neutral to the implementation of resolution. As a corollary, phyloreferences adapt seamlessly to changes in phylogenetic knowledge.

To accomplish the goals of developing a phyloreference specification that meets these requirements and demonstrating its practical utility, we designed 4 major deliverables described in the following sections.

I. Specification for encoding phyloreferences and phylogenies in OWL

A central premise of our proposal is that a phyloreference specification can be developed in an OWL2 framework such that the 4 key requirements laid out above are met. Specifically, our hypothesis is that modeling phyloreferences as class expressions in the OWL2-DL subset, in combination with representing phylogenies in OWL2-DL in a compatible way, can achieve these requirements. In this framework, the process of resolving a phyloreference equates to running a reasoner, a software tool, to classify an ontology consisting of the phyloreference(s) and the phylogeny. Classification of an ontology yields primarily the inferred subclass hierarchy between classes, and the inferred types (class memberships) for instances. Both the subclass hierarchy and type inference includes class expressions. Thus, if the elements of the phylogeny are all modeled as instances of classes from an appropriate ontology, a reasoner can resolve a phyloreference by inferring which elements of the phylogeny are instances of the defining class expression. In principle, the same can be achieved through subclass reasoning if the phylogeny elements are modeled as classes.

In this model, phyloreferences have the following properties. (1) By adding appropriate class or property restrictions to a phyloreference's defining class expression, the elements of a phylogeny it resolves to can be tailored in any way supported by how the representation of the phylogeny is modeled in OWL. This meets requirement 1. (2) The classes, properties, and object values used in class and property restrictions all have globally unique identifiers, normally HTTP URIs. Although it is possible to use data properties in class expressions, whose value would be a literal without identifier, such use is evident to and distinguishable by a machine, and could thus be enforceably disallowed in a specification. A mapping may be required between the (property, class, or instance) identifiers used in a phyloreference and those used in a phylogeny. Such a mapping would have to be in the form of OWL axioms to be available to the reasoner. Hence, requirement 2 is met. (3) The semantics of a phyloreference are necessarily and sufficiently defined through its defining class expression, which specifies which instances are and are not members of the class. These semantics are fully accessible to a reasoner, and if conditions are added, removed, or modified, a reasoner can infer the appropriate subsumption relationship between old and new definition. This meets requirement 3. (4) As a class expression, a phyloreference does not point to specific elements in a specific phylogeny as its matching instances, but instead defines a set of independent conditions, which requires a reasoner to determine whether a given phylogeny has any elements that match. The development of efficient OWL reasoners is a very active area of research, and as a result different reasoners exist for OWL2-DL [122-125] and subprofiles [126] such as OWL2-EL [127]. Any of those with sufficient expressivity can be used for resolving phyloreferences. Therefore, requirement 4 is met as well.

There are numerous possibilities for representing one and the same phylogenetic definition as a phyloreference, and yet more for modeling phylogenies in OWL. The task of specification is to identify, test, and document a single set of modeling patterns that is sufficiently expressive yet computationally efficient. Part of developing the specification will be the development of a shared phyloreferencing ontology containing classes and properties required to follow the modeling patterns. We will focus the specification on the 3 principal types of phylogenetic definitions that have emerged from the extensive work on phylogenetic taxonomy (see Background).

As promising as the premises of doing this in an OWL2 framework are, identifying a model that accommodates these definitions is non-trivial. In particular, the notions of most and least inclusive, or first and most recent ancestor, are challenging to model in OWL; concepts such as biggest or smallest subclass, or closest superclass, are absent from the language. However, as part of prior results, co-PI Lapp has developed, using only off-the-shelf OWL ontology editing

and reasoning tools, an initial proof-of-principle that demonstrates how phyloreferences for all 3 principal types and phylogenies can be encoded in OWL2 such that a standard OWL-DL reasoner can subsequently infer the elements matching the phyloreference [128]. For this to work, a novel way of asserting a phylogeny's topology had to be developed. The model still retains inference of a full set of parent, child, ancestor, and descendant relationships declared in Comparative Data Analysis Ontology (CDAO) [129]. This work is only at the very beginning. Aside from rigorous evaluation, in particular the following questions require considerable further research.

(i) *Consistency of reference.* There are obvious advantages to consuming applications if phyloreferences can be expected to be consistent in what they reference. For example, by convention phyloreferences could always reference the ancestor node, which can then be easily expanded at query (i.e., resolution) time to match the whole clade descending from it. It is not obvious which of the multiple possibilities for reference type (ancestor node, clade descending from ancestor node, branch subtending ancestor node) is advantageous over others. Moreover, the 3 types of phylogenetic definitions differ in what they originate with. For example, branch-based definitions in principle include the branch subtending the ancestor node, and thus for a phylogeny any data and metadata hanging off of that branch.

(ii) *Modeling and identification of node specifiers.* The unambiguity of phyloreferences, as in fact of any formalism for phylogenetic definitions, relies in part on the ability to unambiguously identify the nodes and apomorphies used in a definition (a.k.a. the specifiers). The proof-of-principle skirts this issue by simply using the (URI) identifiers of nodes in the OWL-encoded example phylogenies. A general specification needs to include modeling conventions for common types of node branch specifiers, in particular taxa, specimens and molecular sequence accessions for nodes, and apomorphies (morphological characters) for branches. Where canonical URI identifiers do not yet exist (establishing them is outside of our scope), we will use surrogate identifiers (e.g., EOL identifiers for taxa) for examples and demonstrations. For apomorphies, we expect to follow the ontological representation of natural phenotypes developed by the Phenoscape and HAO projects [112,130,131]. To create corresponding ontological expressions for apomorphies within the angiosperms (flowering plants), we will take advantage of (and where necessary contribute to) the Plant Ontology [132-134] and related rich community ontology resources [135].

(iii) *Scalability and computational efficiency.* A key factor determining scalability of our approach will be reasoner performance, which in turn depends strongly on axiomatic expressivity. OWL2 establishes expressivity profiles [126] to facilitate development of substantially faster reasoners in return for sacrificing some of the full OWL2-DL expressivity. In particular, the ELK reasoner [127], which supports the OWL2-EL profile, has gained prominence as the only reasoner able to classify large and complex ontologies (such as Uberon [107]) in reasonable time. As a consequence, the models for encoding phyloreferences and phylogenies may need to be limited from OWL2-DL to OWL2-EL. Also, experience from the Phenoscape project shows that reasoning with only the Tbox of an ontology (classes and the axioms about them) can be orders of magnitude faster than when the Abox (instances and axioms about them, thus including type inference for instances) is included. Therefore, the model for encoding phylogenies employed in the proof-of-principle may need to be modified such that phyloreference resolution can be cast as subclass reasoning rather than instance type inference.

(iv) *Comprehensive evaluation with more varied phylogenies.* The proof-of-principle so far includes only rooted and strictly bifurcating trees. The models for encoding phyloreferences and phylogenies need to be evaluated for correctness for unrooted trees, polytomies, and for anastomosing phylogenies.

We anticipate this deliverable to result in the following major products. (1) A Phyloreferencing Ontology in OWL2, defining classes, properties, and other entities needed to create phyloreference expressions, and to encode phylogenies in OWL2 so they are suitable for reasoning with phyloreferences. As already demonstrated in the proof-of-principle, this ontology wherever possible will make extensive use of existing ontologies and vocabularies, in particular the Comparative Data Analysis Ontology (CDAO) [129], the Taxon Name Resolution Service ontology [136], and the Darwin Core vocabulary [137]. (2) A written specification, including phyloreference templates and examples, for constructing phyloreferences in OWL2 for phylogenetic definitions of all major types, and for encoding phylogenies in OWL, such that phyloreference resolution is as efficient as possible. (3) A software tool that takes a phylogeny in a standard format as input, and produces an OWL ontology representation that adheres to our specification.

II. OWL ontology of vetted phyloreferences

One of the governing principles of our research plan is that it be guided by use-cases, in the form of phylogenetic definitions that biologists have already shown to need. For our purposes, we define this as phylogenetic definitions published in the peer-reviewed literature. PI Cellinese has already extracted an initial set of 77 clade definitions from published articles in the course of developing and initially populating RegNum, a structured relational database and web-application currently in testing [39]. Although RegNum is designated as the future registration database for the *PhyloCode* [67], it allows the submission of clade definitions unassociated with, or not even compliant with the *PhyloCode* (such as phylogenetically defined species [60]). As a resource that is set to grow with the upcoming official ratification of the *PhyloCode*, this presents an excellent source for ground-proofing our specification and demonstration work. We will therefore develop an automatic software tool that transforms RegNum records into an OWL ontology of named phyloreferences that adhere to our specification. The transformation will retain as many metadata as possible, such as full textual description, attribution (publication, authors) and provenance (associated published phylogeny) information. The resulting OWL ontology will be regularly updated with new records in RegNum (or obviously when the phyloreference specification changes), and will be published online. This will not only make the RegNum content available for wider reuse, but will also make the semantics of its clade definitions fully accessible to machines. As a side effect, it may help to incentivize authors publishing articles with phylogenetic definitions to also submit their definitions to this registry.

The set of definitions prepopulated in RegNum is not as broadly diverse as would be desirable for helping to guide the specification work. It is also not representative enough of the published work on angiosperms to serve well for our demonstration application (see III.). To establish a set that better fits these purposes, we will supplement RegNum's content by curating phylogenetic definitions harvested from the peer-reviewed literature. Angiosperms, or flowering plants, represent one of the largest clades in the Tree of Life, consisting of an estimated >350,000 known species [138], and many more still believed to await discovery. Due to its astonishing diversity, evolutionary patterns, and economic importance, the group has attracted much attention by scientists. As a result, the body of knowledge about angiosperm phylogeny is considerable and steadily increasing, which in turn has motivated many systematists to generate and publish a variety of phylogenetically defined names.

The RegNum web-application already has an interactive user interface for entering clade definitions in a structured form. However, the curation work will necessitate a few additions to this interface, which we will make as part of this project. Specifically, RegNum's interface does not yet support entering ontology-based apomorphic phenotypes. Preliminary work on enabling this capability has already taken place in PI Cellinese's lab, but due to lack of funding has not progressed to a functional interface component. We will base design and implementation of the

interface on the pioneering work done for very similar features of the Phenex phenotype curation tool [108] and the Hymenoptera Anatomy Ontology (HAO) project [130].

III. Proof-of-concept application for utility and correctness of phyloreferences

As a proof-of-concept for the properties and potential of our phyloreferencing model, we will develop an online demonstration application in two phases. In the first phase, described here, we aim to demonstrate correctness and scalability of our phyloreferencing model, as well as the utility of using phyloreferences to query a phylogeny for a large and diverse group. In the second phase, described in the next section (see IV.), we generalize this to demonstrating navigation of a well-known large-scale biodiversity data resource at the scale of the entire Tree of Life.

For the first phase, we will develop an online web-application that will allow users to query a comprehensive phylogenetic tree of angiosperms (flowering plants) by phyloreferences. The tree will be obtained from the Open Tree of Life project (see letter of commitment by K. Cranston) as a synthesis of phylogenetic trees published for angiosperms to date, with missing tip nodes grafted from appropriate taxonomies. We choose this synthesized tree rather than any individually published phylogeny to achieve the best possible coverage for the range of taxon specifiers used in the ontology of angiosperm phyloreferences created in the previous step from published angiosperm clade definitions (see III.). Therefore, the tree against which these phyloreferences will be resolved may differ in topology to the trees in reference to which authors originally stated their phylogenetic definitions. As a domain expert on angiosperm systematics, PI Cellinese, together with the graduate student, will be responsible for evaluating the results of phyloreference resolution for correctness, and she will assess the deviations, if any, between the results and the clades implied by the originally referenced phylogenies.

The user interface of the demonstration application will allow a user to select angiosperm phyloreferences from the ontology of vetted phyloreferences (see section III.), and obtain the matching ancestor node and clade from the tree. A user will also be able to interactively compose a phyloreference on the fly. The result of resolution will be shown graphically, and the matching clade will also be available for download in standard phylogeny formats. We expect that visualizing the tree and query results will reuse an existing browser-side phylogeny visualization tool, such as the treelib-js library first developed by Rod Page and subsequently extended by Ben Morris [139] for use within the PhyloCommons application [140]. We will also enable the inverse query, allowing a user to select nodes on the tree and ask which phyloreferences in the ontology include them in the clades they define. For the server-side backend that executes the queries, we will experiment with two principle approaches. One will be backed by an RDF triple store (run by Virtuoso [141] or Bigdata® [142]) with all ontologies and the OWL-encoded phylogeny pre-reasoned (the reasoning supported by RDF triple stores is not expressive enough for our purposes). This approach is similar to the one that currently underpins the all-vertebrate version of the Phenoscape Knowledgebase [144]. Due to most of the reasoning taking place offline, the efficiency constraints for the reasoning stage are relaxed, and the query language (SPARQL [143] built into the triple store allows very flexible queries and can even provide a generic out-of-the-box API. The downside is that instead of resolving phyloreferences with an OWL reasoned, those entered on-the-fly would essentially need to be translated into triple store queries, which can be error prone in terms of correctly matching OWL semantics, and can sometimes perform poorly.

As an alternative approach, we will explore backing query execution by an OWL reasoner process kept online that holds the ontologies and OWL-encoded phylogeny in memory. Many reasoners, including ELK, can efficiently answer DL queries after initial classification. The effectiveness of coupling this approach with an RDF triple store has already been shown in the Phenoscape project, resulting in a generic tool on which we can readily build [145]. Even if

resolving a phyloreference composed on the fly requires re-classification of the ontology, this is usually possible in a much shorter time than the full initial classification. The high-performance computing (HPC) facilities at the University of Florida [146], where PI Cellinese is based, would allow us to host the application in a way that has direct access to high-memory HPC nodes (as would hosting it on Amazon's EC2 infrastructure [147], and hence such a setup is not constrained to specialized HPC facilities).

IV. Proof-of-concept application for navigating large-scale data resources

In the second phase of developing our demonstration application (see III. for the first), we scale up the phylogeny considerably further to now consist of the full synthetic Open Tree of Life [35]. This tree is currently estimated to comprise approximately 2.5 million tips, which include all 3 major domains of life and viruses. It is the most comprehensive synthesis of Life to date and an ideal platform to test a number of evolutionary hypotheses from large-scale diversification events to processes that have shaped biogeographic patterns. The tree will be obtained in bulk from the Open Tree of Life project (see letter of commitment, K. Cranston).

As another extension of the first phase, we aim to demonstrate how the results of querying the tree, i.e., phyloreference resolution, can be used to navigate even large organism-linked data resources. For this we choose the Encyclopedia of Life (EOL) [34], a well-known biodiversity data aggregator with species-linked data across the diversity of life. The main challenges we anticipate for this part of the demonstration are rooted in the absence of a mapping between the taxonomy used by EOL to organize its data, and the internal nodes in the Open Tree, and thus clade ancestor nodes resulting from phyloreference resolution. However, this problem will likely be common for most organism-linked data resources wishing to employ our framework, and therefore we designed the proof-of-concept to require the development of approaches to address it.

We will develop and demonstrate approaches for two different aspects of this integration problem. In the first, we consider the problem of sub-setting the data of the targeted resource (here, EOL) by the tip nodes of the clade (in this case species-level) to which a phyloreference resolves. This does not require nor involve a backbone hierarchy; even though EOL does use one, many other organism-linked resources do not. Sub-setting by a list of tip nodes included in a clade also ensures that the result is equal to or a subset of the clade as defined by applying a phyloreference to the tree. This approach does require a mapping between the tips of the tree, which here is the synthetic Open Tree of Life, and the names used by the targeted resource, in this case EOL. To create this mapping, we will match names from the Open Tree Taxonomy [148] to the taxonomy sources used by both EOL and Open Tree, for which Open Tree provides an API through its Taxonomic Name Resolution Service (TNRS) [149]. We will then transform the mapping into axiomatic form in OWL2 so that it is available to OWL reasoners. In the demonstration application that we will develop a user, upon executing a phyloreference query, can choose to be transferred to a collection of species pages at EOL that correspond to the tip nodes of the resulting clade from the synthetic Open Tree of Life.

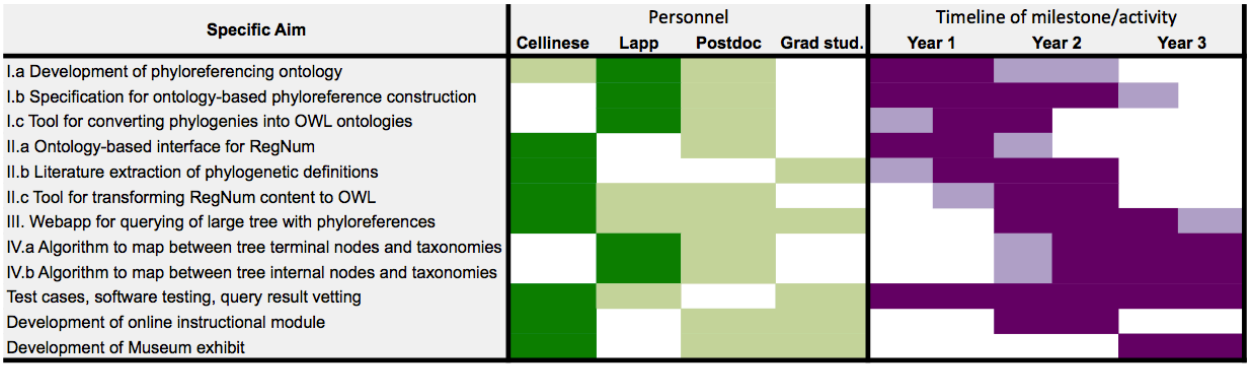
In the second aspect, instead of sub-setting by the set of tip nodes that match a phyloreference we consider the problem of linking to the targeted resource (here, EOL) by an internal node in the taxonomic hierarchy it uses to index its data. Because taxonomies have lower resolution than a phylogeny, the group of organisms subsumed by the best matching taxonomy node need not be a strict subset of the clade implied by the internal node in the phylogeny matching a phyloreference. We will develop a tool to create mappings, in the form of OWL axioms, between internal nodes in a phylogeny and those in a taxonomic hierarchy, and we will develop informative ways to summarize to the user the deviation between the clade implied by the phylogeny node and the best matching one in the taxonomy. These tools will then allow us to extend the previous demonstration application to enable a user to choose to be transferred to

EOL's page for the particular higher-level node of its taxonomy, rather than the collection of species pages.

As a side product this work will result in a mechanism to label internal nodes of the Open Tree based on published phylogenetic definitions. The Open Tree project has expressed strong interest in such a mechanism, and once validated will work to integrate it into their node labeling workflows (letter of commitment, K. Cranston).

MANAGEMENT PLAN

A. Responsibilities and Timelines.



Dark shade: primary responsibility or activity. Light shade: secondary responsibility or activity

The project team includes PI Cellinese and co-PI Lapp, as well as a postdoctoral researcher and a graduate student, both of whom will be based in Cellinese's lab at the Florida Museum of Natural History (FLMNH). PI Cellinese is a domain expert on angiosperm evolution, systematics, and biological nomenclature. She also leads the development of several informatics resources for the phylogenetics, systematics, and biodiversity science communities. She will oversee the project in its entirety, and have responsibility for training and supervision of the postdoc and graduate student. In regard to project deliverables, her responsibilities include steering design and implementation of all software so that outcomes meet project and scientific goals; devising suitable test cases and evaluating correctness of phyloreference query resolution; and directing the delivery of the broader impact activities. She will also develop the online course module about phylogenetic definitions. Co-PI Lapp has extensive expertise in evolutionary bioinformatics, technologies for data integration, and on using semantic web technologies to make the semantics of data amenable to computation. He will co-steer the design, testing, and implementation of software products, and will have primary responsibility for all work related to the development of ontologies and reasoning infrastructure, including development, testing and documentation of the phylorerence specification. In addition, Lapp will ensure that the project's work is properly coordinated with relevant ontology and standards projects. This includes training project personnel in contributing to relevant community ontologies as needed to serve project goals. The postdoctoral researcher will be charged with all aspects of development of the project's software products, including research for performance optimization of reasoning and exploring different backend approaches for the demonstration application. Together with co-PI Lapp, he/she will coordinate with the EOL and Open Tree of Life projects to obtain data and to communicate project requirements and results. He/she will also be involved in the training of the graduate student, and contribute to the broader impact deliverables. The graduate student will search the literature for previously or newly published phylogenetic definitions and will be charged with entering them into the RegNum database. He/she will contribute to generating test cases for phyloreference resolution and to evaluating resolution and query results. Additionally, he/she will contribute to the development of the Museum public exhibit and the on-line course module.

B. Advisory Board

We have assembled an Advisory Board comprised of a diverse group of experts that will meet with the project team face-to-face once every year. The Board's role includes ensuring that the project makes effective use of relevant emerging technologies, and that project outcomes stand to have a wide impact on science beyond the PIs' areas of expertise. Individuals who have agreed to serve on the Board (see letters of commitment), and their areas of expertise, are Jim Balhoff (NESCent and Phenoscape), large-scale reasoning with a querying of ontology-based data; David Baum (University of Wisconsin, Madison), phylogenetics and phylogenetic nomenclature; Holly Bik (University of California, Davis), microbial phylogenomics and tree visualization; Christopher Mungall (Lawrence Berkeley National Laboratory), OWL2 semantics and modeling; Susan Perkins (American Museum of Natural History), microbial phylogenetics; and Michael Sanderson (University of Arizona), large-scale phylogenetics and tree visualization. We reserve one additional seat on the Board for recruitment if and when a gap in a further area of expertise has been determined. The annual face-to-face meeting with the Board will be held at the iDigBio headquarters at the University of Florida (letter of commitment, L. Page), in conjunction with one of the 3 annual project all-hands meetings. This will also provide an opportunity to reach out to the biological collections community, for which iDigBio develops critical cyberinfrastrure for specimen identification and data management.

C. Project Coordination Plan

The physical proximity between postdoc and graduate student with PI Cellinese will facilitate close face-to-face collaboration between them. To ensure effective coordination and communication of work progress, obstacles, and project priorities with co-PI Lapp, who is based at Duke University and shares mentoring and training responsibilities, we will conduct biweekly virtual team meetings. The whole team will also meet 3 times a year face-to-face at the University of Florida, with a focus on reviewing and planning delivery milestones, resolving research and development obstacles, and on critical evaluation of project products. In addition, to facilitate thorough training of the postdoc in scientific software development and the application of ontologies, every year the postdoc will spend a full week at NESCent working closely with co-PI Lapp and participating in knowledge exchange with others in Lapp's Informatics team, which at any given time work on a variety of scientific data management, integration, and semantics problems. Finally, both Cellinese and Lapp have ample experience with electronic communication and remote collaboration in successful distributed project teams, and will bring corresponding best practices to bear on this project.

D. Collaboration and coordination with other projects

Both PIs Cellinese and Lapp will use their existing involvements in cyberinfrastructure and community standards initiatives to coordinate development and project direction with the most relevant ones. In addition to projects mentioned earlier, these include the following. **iDigBio** [150] is a national effort aimed at digitizing all US scientific collections and making them publicly available via a national portal. The **Global Names Architecture** (GNA [151]) aims at aggregating and resolving incongruence among the many idiosyncratic taxon names and descriptive concepts that have been published since Linnaeus. The **BiSciCol** [118] project, for which Cellinese is the PI, aims at identifying and tracking biological collections and all of their derivatives, including specimens and taxon names. Additionally, the BiSciCol team is leading an effort to evaluate current practices in the use of globally unique identifiers (GUIDs) assigned to digital objects. PI Cellinese and collaborators have an article currently in review on this important topic and are organizing a GUIDs pre-TDWG meeting workshop in October 2014 that all major stakeholders will attend, including the OpenTree of Life project. **Phenoscape** [109], for which Lapp is co-PI, transforms natural language descriptions of evolutionary phenotypes into ontological expressions with fully computable semantics, and integrates these with mutant gene phenotypes to enable knowledge discovery. The **Phylogenetics Standards Interest Group** of

TDWG [152] reaches out to biodiversity information scientists. The Minimum Information for a Phylogenetic Analysis (**MIAPA**) [153] standard aims to define the metadata attributes required to make a phylogeny reusable, including common identifier schemes for tip nodes.

E. Dissemination Plan

All software source code and ontologies developed as part of this project will be available on public version control repositories, in particular Github, from the start of development under OSI-compliant open-source and Creative Commons Attribution (CC-BY) licenses, respectively. In addition to traditional peer-reviewed journal publications, project results will be disseminated early on at relevant scientific meetings (e.g., iEvoBio and Evolution Meetings), including domain-specific conferences (such as the Annual Botany Meeting).

BROADER IMPACTS

Educational Impact: This PIs will formally train a postdoc and a graduate student in all aspects of project implementation, from generating deliverables to dissemination of results and broader impact activities. Both postdocs and graduate students will directly interact with collaborators and domain experts. Phylogenetic nomenclature is not formally taught in most universities, likely because the topic is somewhat controversial and readily usable in-depth instructional material is completely lacking. Yet, an increasing number of phylogenetic definitions are being published across multiple domains (see Significance). To promote the wider teaching of this subject, PI Cellinese will develop, in consultation with Museum educational experts, suitable material that will be offered as an independent online module. The content of the module will be based on new lessons she has integrated into courses she currently teaches, such as Biological Nomenclature (grad and undergrad levels), Principle of Systematic Biology (grad level), Evolutionary Biogeography (grad level) and Plant Geography (undergrad level). The module will center on the theoretical underpinnings and practical challenges revolving around phylogenetic nomenclature. Specific topics will include 1) Philosophy and basic principles; 2) Clade names and how to formally construct phylogenetic definitions; 3) The nature of 'species' and needs and consequences for species nomenclature in view of tree-thinking; 4) Genomics, the era of 'Dark taxa', and how to handle phylotypes; 5) The potential of ontologies in evolutionary biology. Every lecture will be accompanied by a number of practical exercises that PI Cellinese has already tested in class settings. Both the postdoc and graduate student will be directly involved in the development of the instructional online module and in teaching the relevant classes as part of the courses listed above. The online module will be publicly available under a CC-BY license through the Museum website (letter of commitment, MacMahon).

Society Impact: The FLMNH is Florida's state museum of natural history and has access to very diverse and broad audiences. The museum will devote a highly visible exhibition space, called the Galleria project, to an exhibit designed by our project team (letter of commitment, MacMahon). The Galleria Project is devoted to communicating the relevance of science to people's daily lives, and showcasing research by museum curators. Under the direction of PI Cellinese, we will create an exhibit that ties together basic concepts in evolutionary biology, the significance of the Tree of Life, and what the ability means to perform queries driven by evolutionary history. The user interface we develop to query the Open Tree of Life will be part of the exhibit in the form of a touchscreen monitor. The FLMNH Museum already has the infrastructure needed to build our proposed exhibit and PI Cellinese has experience in designing exhibits that are easily accessible to the public. The FLMNH is currently showcasing an exhibit on evolution in the Mediterranean basin that she designed with one her graduate student and an entire class of Plant Geography undergraduates. Over one year alone, the FLMNH exhibit will be seen by over 200,000 visitors. Aside from the public outreach, this activity will also allow the graduate student and postdoc to learn about communicating research findings to a general audience.

References

1. OWL 2 Web Ontology Language Document Overview. Available from: http://www.w3.org/TR/owl2-overview/
2. OWL 2 Web Ontology Language Direct Semantics. Available from: http://www.w3.org/TR/owl2-direct-semantics/
3. Campillos, M., Kuhn, M., Gavin, A.-C., Jensen, L.J., and Bork, P. (2008). Drug target identification using side-effect similarity. Science (New York, N.Y.) *321*, 263-266. 10.1126/science.1158140
4. Washington, N.L., Haendel, M.A., Mungall, C.J., Ashburner, M., Westerfield, M., and Lewis, S.E. (2009). Linking human diseases to animal models using ontology-based phenotype annotation. PLoS Biology *7*, e1000247. 10.1371/journal.pbio.1000247
5. Jensen, L.J., and Bork, P. (2010). Ontologies in quantitative biology: a basis for comparison, integration, and discovery. PLoS Biology *8*, e1000374. 10.1371/journal.pbio.1000374
6. Mabee, P.M., Ashburner, M., Cronk, Q., Gkoutos, G.V., Haendel, M., Segerdell, E., Mungall, C., and Westerfield, M. (2007). Phenotype ontologies: the bridge between genomics and evolution. Trends in Ecology & Evolution *22*, 345-350. 10.1016/j.tree.2007.03.013
7. Dahdul, W.M., Lundberg, J.G., Midford, P.E., Balhoff, J.P., Lapp, H., Vision, T.J., Haendel, M.A., Westerfield, M., and Mabee, P.M. (2010). The teleost anatomy ontology: anatomical representation for the genomics age. Systematic Biology *59*, 369-383. 10.1093/sysbio/syq013
8. Mabee, P., Balhoff, J.P., Dahdul, W.M., Lapp, H., Midford, P.E., Vision, T.J., and Westerfield, M. (2012). 500,000 fish phenotypes: The new informatics landscape for evolutionary and developmental biology of the vertebrate skeleton. Journal of Applied Ichthyology *28*, 300-305. 10.1111/j.1439-0426.2012.01985.x
9. Deans, A.R., Yoder, M.J., and Balhoff, J.P. (2011). Time to change how we describe biodiversity. Trends in Ecology & Evolution *27*, 78-84. 10.1016/j.tree.2011.11.007
10. de Queiroz, K., and Gauthier, J. (1990). Phylogeny as a Central Principle in Taxonomy: Phylogenetic Definitions of Taxon Names. Systematic Zoology *39*, 307-322
11. de Queiroz, K., and Gauthier, J. (1992). Phylogenetic Taxonomy. Annual Review of Ecology and Systematics *23*, 449-480. doi:10.1146/annurev.es.23.110192.002313
12. Gauthier, J. (1986). Saurischian monophyly and the origin of birds. In The origin of birds and the evolution of flight, K. Padian, ed. (San Francisco: California Academy of Sciences), pp. 1–55
13. Estes, R., de Queiroz, K., and Gauthier, J. (1988). Phylogenetic relationships within Squamata. In Phylogenetic relationships of the lizard families: essays commemorating Charles L. Camp., R. Estes and G.K. Pregill, eds. (Stanford, California: Stanford University Press), pp. 119–281.
14. Keesey, T.M. (2007). A mathematical approach to defining clade names, with potential applications to computer storage and processing. Zoologica Scripta *36*, 607-621. doi:10.1111/j.1463-6409.2007.00302.x
15. Sereno, P.C. (2005). The Logical Basis of Phylogenetic Taxonomy. Systematic Biology *54*, 595-619
16. Phyloreference - iPToL - iPlant Collaborative Wiki. Available from: https://pods.iplantcollaborative.org/wiki/display/iptol/Phyloreference
17. Phyloinformatics VoCamp - Phyloreferencing Subgroup. Available from: http://www.evoio.org/wiki/Phyloreferencing_subgroup

18. Pesquita, C., Faria, D., Falcão, A.O., Lord, P., and Couto, F.M. (2009). Semantic similarity in biomedical ontologies. PLoS Computational Biology *5*, e1000443. 10.1371/journal.pcbi.1000443
19. Vision, T., Blake, J., Lapp, H., Mabee, P., and Westerfield, M. (2011). Similarity between semantic description sets: addressing needs beyond data integration. In Proceedings of the First International Workshop on Linked Science (LISC 2011), Volume 783, T. Kauppinen, L.C. Pouchard and C. Keßler, eds. (Bonn, Germany: CEUR Workshop Proceedings).
20. Bauer, S., Köhler, S., Schulz, M.H., and Robinson, P.N. (2012). Bayesian ontology querying for accurate and noise-tolerant semantic searches. Bioinformatics (Oxford, England) *28*, 2502-2508. 10.1093/bioinformatics/bts471
21. Schönland, S. (1889). Campanulaceae. In Die Natürlichen Planzenfamilien, Volume 5, A. Engler and K. Prantl, eds. (Leipzig: Engelmann), pp. 40-70.
22. Takhtajan, A. (1980). Outline of the classification of flowering plants (Magnoliophyta). Botanical Review *46*, 225-359
23. Shetler, S.G., and Morin, N.R. (1986). Seed Morphology in North American Campanulaceae. Annals of the Missouri Botanical Garden *73*, 653-688
24. Kolakovskii, A.A. (1994). The conspectus of the system of the Old World Campanulaceae. Botanicheskii Zhurnal (St. Petersburg) *79*, 109-124
25. Takhtajan, A. (1997). Diversity and classification of flowering plants, (New York: Columbia University Press).
26. Burnett, G.T. (1835). Campanulidae. In Outlines in Botany. (London: Churchill), pp. 942, 1094, 1110.
27. Haberle, R.C., Dang, A., Lee, T., Penaflor, C., Cortes-Burn, H., Oestreich, A., Raubenson, L., Cellinese, N., Edwards, E.J., Kim, S.-T., Eddie, W.M.M., and Jansen, R.K. (2009). Taxonomic and biogeographic implications of a phylogenetic analysis of the Campanulaceae based on three chloroplast genes. Taxon *58*, 715-734
28. Knox, E.B. (2014). The dynamic history of plastid genomes in the Campanulaceae sensu lato is unique among angiosperms. Proceedings of the National Academy of Sciences *111*, 11097-11102. 10.1073/pnas.1403363111
29. Wang, Q., Zhou, S.-L., and Hong, D.-Y. (2013). Molecular phylogeny of the platycodonoid group (Campanulaceae s. str.) with special reference to the circumscription of Codonopsis. Taxon *62*, 498-504. 10.12705/623.2
30. Mansion, G., Parolly, G., Crowl, A.A., Mavrodiev, E., Cellinese, N., Oganesian, M., Fraunhofer, K., Kamari, G., Phitos, D., Haberle, R., Akaydin, G., Ikinci, N., Raus, T., and Borsch, T. (2012). How to handle speciose clades? Mass taxon-sampling as a strategy towards illuminating the natural history of Campanula (Campanuloideae). PLoS One *7*, e50076. 10.1371/journal.pone.0050076
31. Crowl, A.A., Mavrodiev, E., Mansion, G., Haberle, R., Pistarino, A., Kamari, G., Phitos, D., Borsch, T., and Cellinese, N. (2014). Phylogeny of Campanuloideae (Campanulaceae) with Emphasis on the Utility of Nuclear Pentatricopeptide Repeat (PPR) Genes. PLoS ONE *9*, e94199. 10.1371/journal.pone.0094199
32. Wang, Q., Wang, X.-Q., Sun, H., Yu, Y., He, X.-J., and Hong, D.-Y. (2014). Evolution of the platycodonoid group (Campanulaceae s. str.) with particular references to biogeography and character evolution. Journal of Integrative Plant Biology, n/a-n/a. 10.1111/jipb.12203
33. Global Biodiversity Information Facility. Available from: http://www.gbif.org
34. Encyclopedia of Life. Available from: http://www.eol.org
35. Open Tree of Life project. Available from: http://opentreeoflife.org/

36. Lin, C.H., Tsai, K.C., Prior, P., and Wang, J.F. (2014). Phylogenetic relationships and population structure of Ralstonia solanacearum isolated from diverse origins in Taiwan. Plant Pathology, n/a-n/a. 10.1111/ppa.12209
37. Page, R.D.M. (2011). Dark taxa: GenBank in a post-taxonomic world, http://iphylo.blogspot.ca/2011/04/dark-taxa-genbank-in-post-taxonomic.html. In iPhylo, Volume 2014.
38. Parr, C.S., Guralnick, R., Cellinese, N., and Page, R.D. (2012). Evolutionary informatics: unifying knowledge about the diversity of life. Trends Ecol Evol *27*, 94-103. 10.1016/j.tree.2011.11.001
39. RegNum - The international clade names repository. Available from: http://wiki.flmnh.ufl.edu/regnum/
40. Smith, S.A., Beaulieu, J.M., Stamatakis, A., and Donoghue, M.J. (2011). Understanding angiosperm diversification using small and large phylogenetic trees. Am. J. Bot. *98*, 404-414. 10.3732/ajb.1000481
41. Goloboff, P.A., Catalano, S.A., Marcos Mirande, J., Szumik, C.A., Salvador Arias, J., Källersjö, M., and Farris, J.S. (2009). Phylogenetic analysis of 73 060 taxa corroborates major eukaryotic groups. Cladistics *25*, 211-230. 10.1111/j.1096-0031.2009.00255.x
42. Kim, O.-S., Cho, Y.-J., Lee, K., Yoon, S.-H., Kim, M., Na, H., Park, S.-C., Jeon, Y.S., Lee, J.-H., Yi, H., Won, S., and Chun, J. (2012). Introducing EzTaxon-e: a prokaryotic 16S rRNA gene sequence database with phylotypes that represent uncultured species. International Journal of Systematic and Evolutionary Microbiology *62*, 716-721. 10.1099/ijs.0.038075-0
43. Massana, R., DeLong, E.F., and Pedros-Alio, C. (2000). A few cosmopolitan phylotypes dominate planktonic archaeal assemblages in widely different oceanic provinces. Applied and Environmental Microbiology *66*, 1777-1787
44. Borchiellini, C., Chombard, C., Manuel, M., Alivon, E., Vacelet, J., and Boury-Esnault, N. (2004). Molecular phylogeny of Demospongiae: implications for classification and scenarios of character evolution. Molecular Phylogenetics and Evolution *32*, 823-837. http://dx.doi.org/10.1016/j.ympev.2004.02.021
45. Joyce, W.G., Parham, J.F., and Gauthier, J.A. (2004). Developing A Protocol For The Conversion Of Rank-Based Taxon Names To Phylogenetically Defined Clade Names, As Exemplified By Turtles. Journal of Paleontology *78*, 989-1013. 10.1666/0022-3360(2004)078<0989:DAPFTC>2.0.CO;2
46. Cantino, P.D., Doyle, J.A., Graham, S.W., Judd, W.S., Olmstead, R.G., Soltis, D.E., Soltis, P.S., and Donoghue, M.J. (2007). Towards a phylogenetic nomenclature of Tracheophyta. Taxon *56*, 822-846
47. Conrad, J.L., Ast, J.C., Montanari, S., and Norell, M.A. (2011). A combined evidence phylogenetic analysis of Anguimorpha (Reptilia: Squamata). Cladistics *27*, 230-277. 10.1111/j.1096-0031.2010.00330.x
48. Soltis, D.E., Smith, S.A., Cellinese, N., Wurdack, K.J., Tank, D.C., Brockington, S.F., Refulio-Rodriguez, N.F., Walker, J.B., Moore, M.J., Carlsward, B.S., Bell, C.D., Latvis, M., Crawley, S., Black, C., Diouf, D., Xi, Z., Rushworth, C.A., Gitzendanner, M.A., Sytsma, K.J., Qiu, Y.-L., Hilu, K.W., Davis, C.C., Sanderson, M.J., Beaman, R.S., Olmstead, R.G., Judd, W.S., Donoghue, M.J., and Soltis, P.S. (2011). Angiosperm phylogeny: 17 genes, 640 taxa. American Journal of Botany *98*, 704-730. 10.3732/ajb.1000404
49. Cárdenas, P., Pérez, T., and Boury-Esnault, N. (2012). Sponge systematics facing new challenges. In Advances in Sponge Science: Phylogeny, Systematics, Ecology, Volume 61, M.A. Becerro, M.J. Uriz, M. Maldonado and X. Turon, eds. (London: Academic Press), pp. 79-209.

50. Hill, M.S., Hill, A.L., Lopez, J., Peterson, K.J., Pomponi, S., Diaz, M.C., Thacker, R.W., Adamska, M., Boury-Esnault, N., Cárdenas, P., Chaves-Fonnegra, A., Danka, E., De Laine, B.-O., Formica, D., Hajdu, E., Lobo-Hajdu, G., Klontz, S., Morrow, C.C., Patel, J., Picton, B., Pisani, D., Pohlmann, D., Redmond, N.E., Reed, J., Richey, S., Riesgo, A., Rubin, E., Russell, Z., Rützler, K., Sperling, E.A., di Stefano, M., Tarver, J.E., and Collins, A.G. (2013). Reconstruction of Family-Level Phylogenetic Relationships within Demospongiae (Porifera) Using Nuclear Encoded Housekeeping Genes. PLoS ONE *8*, e50437. 10.1371/journal.pone.0050437
51. Mannion, P.D., Upchurch, P., Barnes, R.N., and Mateus, O. (2013). Osteology of the Late Jurassic Portuguese sauropod dinosaur Lusotitan atalaiensis (Macronaria) and the evolutionary history of basal titanosauriforms. Zoological Journal of the Linnean Society *168*, 98-206. 10.1111/zoj.12029
52. Schoch, R.R. (2013). The evolution of major temnospondyl clades: an inclusive phylogenetic analysis. Journal of Systematic Palaeontology *11*, 673-705. 10.1080/14772019.2012.699006
53. Sterli, J., Pol, D., and Laurin, M. (2013). Incorporating phylogenetic uncertainty on phylogeny-based palaeontological dating and the timing of turtle diversification. Cladistics *29*, 233-246. 10.1111/j.1096-0031.2012.00425.x
54. Torres-Carvajal, O., and Mafla-Endara, P. (2013). Evolutionary history of Andean Pholidobolus and Macropholidus (Squamata: Gymnophthalmidae) lizards. Molecular Phylogenetics and Evolution *68*, 212-217. http://dx.doi.org/10.1016/j.ympev.2013.03.013
55. Wojciechowski, M.F. (2013). Towards a new classification of Leguminosae: Naming clades using non-Linnaean phylogenetic nomenclature. South African Journal of Botany *89*, 85-93. http://dx.doi.org/10.1016/j.sajb.2013.06.017
56. Hundt, P.J., Iglésias, S.P., Hoey, A.S., and Simons, A.M. (2014). A multilocus molecular phylogeny of combtooth blennies (Percomorpha: Blennioidei: Blenniidae): Multiple invasions of intertidal habitats. Molecular Phylogenetics and Evolution *70*, 47-56. http://dx.doi.org/10.1016/j.ympev.2013.09.001
57. Rabi, M., Sukhanov, V.B., Egorova, V.N., Danilov, I., and Joyce, W.G. (2014). Osteology, relationships, and ecology of Annemys (Testudines, Eucryptodira) from the Late Jurassic of Shar Teg, Mongolia, and phylogenetic definitions for Xinjiangchelyidae, Sinemydidae, and Macrobaenidae. Journal of Vertebrate Paleontology *34*, 327-352. 10.1080/02724634.2013.807274
58. Hennig, W. (1966). Phylogenetic Systematics, (Urbana: University of Illinois Press).
59. de Queiroz, K. (1988). Systematics and the Darwinian Revolution. Philosophy of Science *55*, 238-259
60. Cellinese, N., Baum, D.A., and Mishler, B.D. (2012). Species and Phylogenetic Nomenclature. Systematic Biology *61*, 885-891. 10.1093/sysbio/sys035
61. Ghiselin, M.T. (1984). "Definition," "Character," and Other Equivocal Terms. Systematic Zoology *33*, 104-110
62. Gauthier, J., and Padian, K. (1985). Phylogenetic, functional, and aerodynamic analyses of the origin of birds and their flight. In The beginnings of birds K. Hecht, G.H. Ostrom, G. Viohl and P. Wellnhofer, eds. (Eichstatt, Germany.: Freude des Jura-Museums), pp. 185–197.
63. Rowe, T. (1987). Definition and Diagnosis in the Phylogenetic System. Systematic Zoology *36*, 208-211
64. de Queiroz, K. (1987). Phylogenetic systematics of iguanine lizards. A comparative osteological study. . Univ. Calif. Publ. Zool. *118*, 1–203
65. Gauthier, J., Estes, R., and de Queiroz, K. (1988). A phylogenetic analysis of Lepidosauromorpha. In Phylogenetic relationships of the lizard families: essays

commemorating Charles L. Camp., R. Estes and G.K. Pregill, eds. (Stanford, California.: Stanford University Press), pp. 15–98.
66. de Queiroz, K., and Gauthier, J. (1994). Toward a phylogenetic system of biological nomenclature. Trends in Ecology & Evolution *9*, 27-31
67. PhyloCode. Available from: http://www.ohio.edu/phylocode/
68. de Queiroz, K. (1992). Phylogenetic definitions and taxonomic philosophy. Biology and Philosophy *7*, 295-313
69. Rowe, T., and Gauthier, J. (1992). Ancestry, paleontology and definition of the name Mammalia. Systematic Biology *41*
70. Bryant, H.N. (1994). Comments on the Phylogenetic Definition of Taxon Names and Conventions Regarding the Naming of Crown Clades. Systematic Biology *43*, 124-130
71. de Queiroz, K. (1994). Replacement of an Essentialistic Perspective on Taxonomic Definitions as Exemplified by the Definition of "Mammalia". Systematic Biology *43*, 497-510
72. Bryant, H.N. (1996). Explicitness, Stability, and Universality in the Phylogenetic Definition and Usage of Taxon Names: A Case Study of the Phylogenetic Taxonomy of the Carnivora (Mammalia). Systematic Biology *45*, 174-189
73. Bryant, H.N. (1997). Cladistic information in phylogenetic definitions and designated phylogenetic contexts for the use of taxon names. Biological Journal of the Linnean Society *62*, 495–503.
74. de Queiroz, K. (1997). The Linnaean hierarchy and the evolutionization of taxonomy, with emphasis on the problem of nomenclature. Aliso, 125-144
75. Sundberg, P., and Pleijel, F. (1994). Phylogenetic classification and the definition of taxon names. Zoologica Scripta *23*, 19–25.
76. Christoffersen, M.L. (1995). Cladistic Taxonomy, Phylogenetic Systematics, and Evolutionary Ranking. Systematic Biology *44*, 440-454
77. Schander, C., and Thollesson, M. (1995). Phylogenetic taxonomy-some comments. Zoologica Scripta *24*, 263-268. doi:10.1111/j.1463-6409.1995.tb00404.x
78. Lee, M.S.Y. (1996). Stability in Meaning and Content of Taxon Names: An Evaluation of Crown-Clade Definitions. Proceedings: Biological Sciences *263*, 1103-1109
79. Lee, M.S.Y. (1996). The phylogenetic approach to biological taxonomy: practical aspects. Zoologica Scripta *25*, 187-190. doi:10.1111/j.1463-6409.1996.tb00159.x
80. Lee, M.S.Y. (1998). Phylogenetic Uncertainty, Molecular Sequences, and the Definition of Taxon Names. Systematic Biology *47*, 719-726
81. Lee, M.S.Y. (1998). Ancestors and taxonomy. Trends in Ecology & Evolution *13*, 26
82. Lee, M.S.Y. (1999). Reference Taxa and Phylogenetic Nomenclature. Taxon *48*, 31-34
83. Lee, M.S.Y. (1999). Stability of higher taxa in phylogenetic nomenclature — some comments on Moore (1998). Zoologica Scripta *28*, 361-366. doi:10.1046/j.1463-6409.1999.00017.x
84. Lee, M.S.Y. (2001). On Recent Arguments for Phylogenetic Nomenclature. Taxon *50*, 175-180
85. Lee, M.S.Y. (2005). Choosing reference taxa in phylogenetic nomenclature. Zoologica Scripta *34*, 329-331
86. Wyss, A.R., and Meng, J. (1996). Application of phylogenetic taxonomy to poorly resolved crown clades: a stem-modified node-based definition of Rodentia. Systematic Biology *45*, 559–568
87. Brochu, C.A. (1997). Synonymy, redundancy, and the name of the crocodile stem-group. Journal of Vertebrate Paleontology *17*, 448–449
88. Cantino, P.D., Olmstead, R.G., and Wagstaff, S.J. (1997). A Comparison of Phylogenetic Nomenclature with the Current System: A Botanical Case Study. Systematic Biology *46*, 313-331

89. Cantino, P.D., Bryant, H.N., Queiroz, K.D., Donoghue, M.J., Eriksson, T., Hillis, D.M., and Lee, M.S.Y. (1999). Species Names in Phylogenetic Nomenclature. Systematic Biology *48*, 790-807
90. Kron, K.A. (1997). Exploring alternative systems of classification. Aliso *15*, 105–111
91. Baum, D.A., Alverson, W.S., and Nyffeler, R. (1998). A durian by any other name: taxonomy and nomenclature of the core Malvales. Harvard Papers of Botany *3*, 315–330
92. Eriksson, T., Donoghue, M.J., and Hibbs, M.S. (1998). Phylogenetic analysis of Potentilla using DNA sequences of nuclear ribosomal internal transcribed spacers (ITS), and implications for the classification of Rosoideae (Rosaceae). Plant Systematics and Evolution *211*, 155-179
93. Härlin, M., and Sundberg, P. (1998). Taxonomy and Philosophy of Names. Biology and Philosophy *13*, 233-244
94. Hibbett, D.S., and Donoghue, M.J. (1998). Integrating Phylogenetic Analysis and Classification in Fungi. Mycologia *90*, 347-356
95. Moore, G. (1998). A Comparison of Traditional and Phylogenetic Nomenclature. Taxon *47*, 561-579
96. Mishler, B.D. (1999). Getting rid of species? In Species: New Interdisciplinary Essays, R. Wilson, ed. (Cambridge: MIT Press), pp. 307-315.
97. Pleijel, F. (1999). Phylogenetic Taxonomy, a Farewell to Species, and a Revision of Heteropodarke (Hesionidae, Polychaeta, Annelida). Systematic Biology *48*, 755-789
98. Sereno, P.C. (1999). Definitions in Phylogenetic Taxonomy: Critique and Rationale. Systematic Biology *48*, 329-351
99. Brochu, C.A., and Sumrall, C.D. (2001). PHYLOGENETIC NOMENCLATURE AND PALEONTOLOGY. Journal of Paleontology *75*, 754-757
100. Bremer, K. (2000). Phylogenetic nomenclature and the new ordinal system of the angiosperms. In Plant systematics for the 21st century, B. Nordenstam, G. El Ghazaly and M. Kassas, eds. (London: Portland Press), pp. 125-133.
101. Judd, W.S., Stern, W., and Cheadle, V.I. (1993). Phylogenetic position of Apostasia and Neuwiedia (Orchidaceae). Botanical Journal of the Linnean Society *113*, 87–94
102. Judd, W.S., Sanders, R.W., and Donoghue, M.J. (1994). Angiosperm family pairs: preliminary phylogenetic analyses. Harvard Papers of Botany *5*, 1–51
103. Alverson, W.S., Whitlock, B.A., Nyffeler, R., Bayer, C., and Baum, D.A. (1999). Phylogeny of the core Malvales: evidence from ndhF sequence data. Am. J. Bot. *86*, 1474-1486
104. Sereno, P., McAllister, S., and Brusatte, S.L. (2005). TaxonSearch: a relational database for suprageneric taxa and phylogenetic definitions. Phyloinformatics *8*, 1-21
105. Mungall, C.J., Gkoutos, G.V., Smith, C.L., Haendel, M.A., Lewis, S.E., and Ashburner, M. (2010). Integrating phenotype ontologies across multiple species. Genome Biology *11*, R2. 10.1186/gb-2010-11-1-r2
106. Mungall, C.J., Bada, M., Berardini, T.Z., Deegan, J., Ireland, A., Harris, M.A., Hill, D.P., and Lomax, J. (2011). Cross-product extensions of the Gene Ontology. Journal of Biomedical Informatics *44*, 80-86. 10.1016/j.jbi.2010.02.002
107. Mungall, C.J., Torniai, C., Gkoutos, G.V., Lewis, S.E., and Haendel, M.a. (2012). Uberon, an integrative multi-species anatomy ontology. Genome Biology *13*, R5. 10.1186/gb-2012-13-1-r5
108. Balhoff, J.P., Dahdul, W.M., Kothari, C.R., Lapp, H., Lundberg, J.G., Mabee, P., Midford, P.E., Westerfield, M., and Vision, T.J. (2010). Phenex: Ontological Annotation of Phenotypic Diversity. PLoS ONE *5*, e10500. 10.1371/journal.pone.0010500
109. Phenoscape project website. Available from: http://phenoscape.org
110. Creative Commons — Attribution 2.0 Generic — CC BY 2.0. Available from: http://creativecommons.org/licenses/by/2.0/

111. Creative Commons Zero 1.0 Universal (CC0 1.0) Public Domain Dedication. Available from: http://creativecommons.org/publicdomain/zero/1.0/
112. Balhoff, J., Midford, P., and Lapp, H. (2011). Integrating anatomy and phenotype ontologies with taxonomic hierarchies. In ICBO: International Conference on Biomedical Ontology (2011) Proceedings, Volume 5. (Buffalo, NY, USA), pp. 8-9.
113. Cellinese, N. (In Press). Campanulaceae. In Phylonyms: A Companion to the PhyloCode, K. de Queiroz, P.D. Cantino and J. Gauthier, eds. (Berkeley: University of California Press).
114. Cellinese, N. (In Press). Campanuloideae. In Phylonyms: A Companion to the PhyloCode, K. de Queiroz, P.D. Cantino and J. Gauthier, eds. (Berkeley, California: University of California Press).
115. Cellinese, N. (In Press). Lobelioideae. In Phylonyms: A Companion to the PhyloCode, K. de Queiroz, P.D. Cantino and J. Gauthier, eds. (Berkeley, California: University of California Press).
116. Beaman, R.S., Traub, G.H., Dell, C.A., Santiago, N., Koh, J., and Cellinese, N. (2012). TOLKIN – Tree of Life Knowledge and Information Network: Filling a Gap for Collaborative Research in Biological Systematics. PLoS ONE *7*, e39352. 10.1371/journal.pone.0039352
117. TOLKIN: Tree of Life Knowledge and Information Network. Available from: http://www.tolkin.org
118. BiSciCol: Biological Science Collections. Available from: http://biscicol.blogspot.com/p/biscicol-prototype.html
119. Stucky, B., Deck, J., Conlin, T., Ziemba, L., Cellinese, N., and Guralnick, R. (2014). The BiSciCol Triplifier: bringing biodiversity data to the Semantic Web. BMC Bioinformatics *15*, 257
120. Cellinese lab Github. Available from: https://github.com/FLMNH-Informatics
121. Phyloinformatics VoCamp. Available from: http://www.evoio.org/wiki/VoCamp1
122. Haarslev, V., and Möller, R. (2003). Racer: A core inference engine for the semantic web. In Proceedings of the 2nd International Workshop on Evaluation of Ontology-based Tools, Volume 87. (Citeseer).
123. Tsarkov, D., and Horrocks, I. (2006). FaCT++ description logic reasoner: System description. In Automated Reasoning (Series: Lecture Notes in Computer Science), U. Furbach and N. Shankar, eds. (Springer), pp. 292-297.
124. Sirin, E., Parsia, B., Grau, B., Kalyanpur, a., and Katz, Y. (2007). Pellet: A practical OWL-DL reasoner. Web Semantics: Science, Services and Agents on the World Wide Web *5*, 51-53. 10.1016/j.websem.2007.03.004
125. Shearer, R., Motik, B., and Horrocks, I. (2008). HermiT: a highly-efficient OWL reasoner. In Proceedings of the 5th International Workshop on OWL: Experiences and Directions (OWLED 2008), C. Dolbear, A. Ruttenberg and U. Sattler, eds. (Karlsruhe, Germany).
126. OWL 2 Web Ontology Language Profiles. Available from: http://www.w3.org/TR/owl2-profiles/
127. Kazakov, Y., Krötzsch, M., and Simančík, F. (2012). ELK reasoner: Architecture and evaluation. In Proceedings of the 1st International Workshop on OWL Reasoner Evaluation (ORE-2012), I. Horrocks, M. Yatskevich and E. Jimenez-Ruiz, eds.
128. Phyloreferencing using OWL Ontologies and Reasoning. Available from: https://github.com/hlapp/phyloref
129. Prosdocimi, F., Chisham, B., Pontelli, E., Thompson, J.D., and Stoltzfus, A. (2009). Initial Implementation of a comparative Data Analysis Ontology. Evolutionary Bioinformatics Online, 47-66

130. Balhoff, J.P., Miko, I., Yoder, M.J., Mullins, P.L., and Deans, A.R. (2013). A semantic model for species description, applied to the ensign wasps (Hymenoptera: Evaniidae) of New Caledonia. Systematic Biology. 10.1093/sysbio/syt028
131. Dahdul, W.M., Balhoff, J.P., Engeman, J., Grande, T., Hilton, E.J., Kothari, C., Lapp, H., Lundberg, J.G., Midford, P.E., Vision, T.J., Westerfield, M., and Mabee, P.M. (2010). Evolutionary Characters, Phenotypes and Ontologies: Curating Data from the Systematic Biology Literature. PLoS ONE *5*, e10708. 10.1371/journal.pone.0010708
132. Avraham, S., Tung, C.W., Ilic, K., Jaiswal, P., Kellogg, E.A., McCouch, S., Pujar, A., Reiser, L., Rhee, S.Y., Sachs, M.M., Schaeffer, M., Stein, L., Stevens, P., Vincent, L., Zapata, F., and Ware, D. (2007). The Plant Ontology Database: a community resource for plant structure and developmental stages controlled vocabulary and annotations. Nucleic Acids Research *36*, D449-D454. 10.1093/nar/gkm908
133. Ilic, K., Kellogg, E.A., Jaiswal, P., Zapata, F., Stevens, P.F., Vincent, L.P., Avraham, S., Reiser, L., Pujar, A., Sachs, M.M., Whitman, N.T., McCouch, S.R., Schaeffer, M.L., Ware, D.H., Stein, L.D., and Rhee, S.Y. (2007). The plant structure ontology, a unified vocabulary of anatomy and morphology of a flowering plant. Plant Physiol *143*, 587-599. 10.1104/pp.106.092825
134. Pujar, A., Jaiswal, P., Kellogg, E.A., Ilic, K., Vincent, L., Avraham, S., Stevens, P., Zapata, F., Reiser, L., Rhee, S.Y., Sachs, M.M., Schaeffer, M., Stein, L., Ware, D., and McCouch, S. (2006). Whole-plant growth stage ontology for angiosperms and its application in plant biology. Plant Physiol *142*, 414-428. 10.1104/pp.106.085720
135. Walls, R.L., Athreya, B., Cooper, L., Elser, J., Gandolfo, M.A., Jaiswal, P., Mungall, C.J., Preece, J., Rensing, S., Smith, B., and Stevenson, D.W. (2012). Ontologies as integrative tools for plant science. American Journal of Botany *99*, 1263-1275. 10.3732/ajb.1200222
136. TNRS Ontology and RDF Data Model. Available from: https://github.com/phylotastic/ontologies/tree/master/tnrs
137. Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., Robertson, T., and Vieglais, D. (2012). Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLoS ONE *7*, e29715. 10.1371/journal.pone.0029715
138. Paton, A.J., Brummitt, N., Govaerts, R., Harman, K., Hinchcliffe, S., Allkin, B., and Lughadha, E.N. (2008). Towards Target 1 of the Global Strategy for Plant Conservation: a working list of all known plant speciesprogress and prospects. Taxon *57*, 602-611
139. treelib-js - Javascript phylogenetic tree library. Available from: https://github.com/bendmorris/treelib-js
140. PhyloCommons, a community phylogenetic database. Available from: http://www.phylocommons.org/
141. OpenLink Virtuoso Universal Server. Available from: http://virtuoso.openlinksw.com/
142. bigdata®. Available from: http://www.systap.com/bigdata.htm
143. SPARQL 1.1 Overview. Available from: http://www.w3.org/TR/sparql11-overview/
144. Phenoscape RDF Knowledgebase. Available from: http://rdf.phenoscape.org/
145. owlet – a query expansion preprocessor for SPARQL. Available from: https://github.com/phenoscape/owlet
146. HPC Facilities at the University of Florida. Available from: http://www.hpc.ufl.edu
147. Amazon Web Services - Elastic Compute Cloud. Available from: http://aws.amazon.com/ec2/
148. Open Tree of Life Reference Taxonomy (OTT). Available from: http://files.opentreeoflife.org/ott/
149. Open Tree of Life APIs. Available from: https://github.com/OpenTreeOfLife/opentree/wiki/Open-Tree-of-Life-APIs
150. iDigBio - Integrated Digitized Biocollections. Available from: http://www.idigbio.org

151. Global Names Architecture. Available from: http://www.globalnames.org
152. Phylogenetics Standards Interest Group (TDWG). Available from: http://www.tdwg.org/activities/phylogenetics/
153. MIAPA - Minimum Information About A Phylogenetic Analysis. Available from: http://www.evoio.org/wiki/MIAPA