Title: *Phyloreferences: Tree-Native, Reproducible, and Machine-Interpretable Taxon Concepts*

Authors: Nico Cellinese(1), Stijn Conix(2), Hilmar Lapp(3)

Affiliations:

(1) Florida Museum of Natural History, University of Florida, Gainesville, FL, USA
(2) Centre for Logic and Philosophy of Science, KU Leuven, Leuven, Belgium
(3) Center for Genomic and Computational Biology, Duke University, Durham, NC, USA

Abstract

Evolutionary and organismal biology, similar to other fields in biology, have become inundated with data. At the same rate, we are experiencing a surge in broader evolutionary and ecological syntheses for which tree-thinking is the staple for a variety of post-tree analyses. To fully take advantage of this wealth of data to discover and understand large-scale evolutionary and ecological patterns, computational data integration, i.e. the use of machines to link data at large scale by shared entities, is crucial. The most common shared entity by which evolutionary and ecological data need to be linked is the taxon to which they belong. In this paper, we propose a set of requirements that a system for defining such taxa should meet for computational data science: taxon definitions should maintain conceptual consistency, be reproducible via a known algorithm, be computationally automatable, and be applicable across the tree of life. We argue that Linnaean names based in Linnaean taxonomy, by far the most prevalent means of linking data to taxa, fail to meet these requirements due to fundamental theoretical and practical shortfalls. We argue that for the purposes of data-integration we should instead use phylogenetic clade definitions transformed into formal logic expressions. We call such expressions *phyloreferences,* and argue that, unlike Linnaean names, they meet all requirements for effective data-integration.

1. Introduction

The last two decades have witnessed a vast increase of available digital biodiversity data. This richness in data has been fostered, in part, by a call to mass-digitize museum repositories (Beaman and Cellinese 2012; Page et al. 2015), and is fueled by the emergence of new applications and data sources, analytical methods, faster algorithms, and improved environmental sensors, among others (Philippe et al. 2005; Porter et al. 2009; Michener and Jones 2012; Chan and Ragan, 2013; Hampton et al. 2017; Kozlov et al. 2019). Additionally, it has led to a corresponding increasing need for digital access, sharing, and re-purposing of data, and, consequently, to a need of using machines to link data from different sources to shared entities. The natural framework for such synthesizing of biodiversity data is the Tree of Life. Tree-thinking has seized a prominent role in systematics since the advent of phylogenetics (Zimmermann 1931, 1934, 1943; Hennig 1950, 1966). The rapidly increasing knowledge across the Tree of Life has now enabled a synthesis of phylogenetic hypotheses on a Tree of Life scale, to produce an encompassing – and digitally fully reusable – view of Life's evolution, the Open Tree of Life (Hinchliff et al. 2015; McTavish

et al. 2017). As a comprehensive and repeatable phylogenetic synthesis, it provides unprecedented opportunities for studying evolutionary patterns across all clades, at large as well as small scales. These clades are the perfect locus at which to integrate the suite of different data types resulting from evolutionary and biodiversity research (e.g., Allen et al. 2018; Eliason et al. 2019; Folk et al. 2019; Howard et al. 2019).

Thus, a system of defining clades is needed to link the vast amount of available biodiversity data in a way that it can be recovered, aggregated, and integrated. However, there is wide disagreement about which system should be used for this purpose. Currently, most biological data and knowledge are directly or indirectly linked to biological taxa via Linnaean taxon names. As we will discuss below, it is well known that in its current shape the Linnaean system leads to numerous problems when applied to data-intensive science that depends on computation. Therefore, an alternative is needed. Broadly speaking, there are two main candidates for such an alternative: to modify the current Linnaean system such that it can fulfill certain requirements (see list below), or, more radically, to abandon the Linnaean system in this context and implement a purely phylogenetic system for clade definitions. The former of these involves repurposing Linnaean names to refer to clades, and using these names as labels for taxon concepts[1]. In that sense, this option is a hybrid between the Linnaean and a phylogenetic system. The latter of these, instead, consists in generating purely phylogenetic definitions of clades.

To arbitrate between these alternatives, we propose the following four requirements that any system suitable for data-integration should meet: (i) The mapping maintains conceptual consistency, meaning, when mapped to different phylogenies, the semantics of the retrieved clades are consistent[2]. (ii) The mapping of a given clade concept to a given phylogenetic hypothesis is exactly reproducible via a known algorithm. (iii) The algorithm to (re)produce the mapping is computationally automatable, which is necessary for processing the very large phylogenies and datasets characteristic of modern biology. This means consulting expert opinion cannot be part of

[1] A taxon concept is the underlying meaning of a group (taxon), whether the group is defined by traits (Linnaean taxonomy) or diagnosed by traits (phylogenetic taxonomy).

[2] By semantics we mean the study, processing, and representation of meaning. The term is used in distinct disciplines, including linguistics and philosophy. In this paper, we use semantics in the sense of computational semantics, which concerns itself with the construction of and automated reasoning with representations of meaning (such as ontologies and logic expressions using ontologies) of natural language expressions.

the algorithm. (iv) The system is applicable to all lineages in the Tree of Life, including in particular those where Linnaean names are not available (e.g., Archaea, fungi, etc.).

In this paper, we show that it is in principle impossible for the Linnaean system to meet these requirements, and present a purely phylogenetic alternative that does meet them. In section 2 we elaborate on the problems of the Linnaean system, and show that it is beyond repair. In section 3 we introduce the purely phylogenetic approach, and show how it can address the shortcomings of the Linnaean system. In section 4 then we introduce one way in which such a phylogenetic alternative could be implemented, namely, *phyloreferences*, and in section 5 we argue that this implementation is preferable over other existing implementations. Finally, in section 6 we address various objections to our proposal, and section 7 concludes the paper.

First of all, it is important to emphasize that the issue at stake in this paper is not that of nomenclature. The question of how to define taxon concepts for data integration is independent from the question of whether these taxon concepts also are named, and even whether these names are Linnaean names. While the approach we propose in this paper fits more naturally with a form of phylogenetic nomenclature, it is also compatible with retaining Linnaean names. Related to this, the issue at stake is not that of whether we should recognize certain taxa as species (Mishler and Wilkins 2018). While a phylogenetic approach like the one proposed here denies that there is an ontological difference between taxa at different levels, it is compatible with recognizing some of these taxa as species. Thus, what is at stake is the best way of defining taxa for data integration, and not the names of these taxa or whether they can be listed as species.

2. The Poverty of Linnaean Names

Many authors before us have pointed to problems caused by Linnaean nomenclature and classification. This section instead discusses two problems of the Linnaean system that make it unsuitable for data integration, and argues that it is not possible to eliminate these problems simply by making small changes to the system.

2.1. The Linnaean Shortfall Limits Data Discovery

A first problem of the Linnaean system is often referred to as the 'Linnaean shortfall'. This is the significant gap in our current knowledge of described vs unknown biological diversity (Brown and Lomolino 1998; Hortal et al. 2015), and highlights our limited ability to first discover and then describe taxa according to the rules of nomenclatural codes. In view of the 6th mass extinction we are currently experiencing (Brook et al. 2008), this represents a true plague in biodiversity science because it implies that we are also losing unknown diversity, and the diversity we do discover is not described (in a Linnaean framework) fast enough. From a computational perspective, the latter point represents a true obstacle to addressing the computable taxon concept challenge because taxa need to be described before they can serve as loci to link data.

Two causes of the Linnaean shortfall are particularly relevant in this context. First, the process of describing diversity is very time consuming and relies on detailed comparative studies of specimens in museum's repositories and field observations. Second, there are far more levels of clades in the Tree of Life than there are ranks to name them. As a result, we continue to discover lineages that have no formal Linnaean names, and for which data can therefore not easily be recovered for reuse. Adopted placeholders such as 'phylotype X' or 'clade A' may serve their purpose within a publication, but are not discoverable and reusable terms (also, see appendix in de Queiroz and Donoghue 2013). This predicament is particularly true in Archaea and other prokaryotes, but very common in many eukaryotes, too, and these lineages have often been referred to as 'dark taxa' (Parr et al. 2012).

The result is that there is a lot of data about taxa that cannot yet, and may never be, linked to Linnaean names. This way, the Linnaean system fails to meet requirement (iv), i.e. to provide the tools to define, communicate and query these unnamed taxa.

2.2. Linnaean names make data discovery difficult to reproduce

One might argue that the rate of species descriptions and formal names could, in principle, increase dramatically and thus alleviate the problem described in the previous subsection. This subsection argues that even if that were the case, Linnaean names would not be suitable for integrating data from different sources. This is because it falls short of the three other requirements as well: (i) it fails to maintain conceptual consistency, (ii) the mapping of a Linnaean name to a phylogeny is not reproducible by a known algorithm, and (iii) the algorithm to do this mapping is not automatable.

To see why the Linnaean system falls short of these requirements, it is helpful to briefly consider its design and history. Prior to Linnaeus, biological knowledge was organized in large, poorly defined categories, and nomenclature was completely unstructured. Linnaeus was a revolutionary for his time, not so much for the system he created (other botanists before him experimented with the ranking system), but for what he enabled. He brought order by formalizing criteria to define logical relationships among abstract classes (categorical ranks) and restructuring the nomenclatural system by enforcing 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. In this context, Linnaean nomenclature served the need of linking names to taxon groups.

Darwinian theory then revolutionized the perspective on biological relationships and taxon group membership, with the notion that it is natural processes that give rise to taxa, while characters can only diagnose, but not define categories (Darwin 1859). Zimmermann (1931, 1934, 1943) and Hennig (1950, 1966) formalized these theories and provided the criteria to construct phylogenetic trees. In this theoretical framework, in which taxa are no longer seen as static entities, it quickly became clear that the phylogeny-governed hierarchy of Hennig's framework is better suited for defining taxa than the logical relatedness of groups in Linnaeus' hierarchical framework (see also Ereshefsky 2001). Consequently, as common practice Linnaean nomenclature has been repurposed to link names to clades. In this hybrid system, Linnaean names are used to label taxon concepts, which are clades rather than fixed entities defined by a set of characters.

However, the Linnaean elements that this hybrid system retains make it impossible to be used for effective data-integration. There are three reasons for this.

First, repurposed Linnaean names define taxon concepts by means of a type specimen and description. However, whenever the type is missing from the phylogeny - which is typically the case - there are no agreed rules for mapping type specimens to clades. Instead, this mapping relies on expert judgement. As different experts tend to do this in different ways (see our example of *Campanula* below), this means that the Linnaean system does not meet requirement (ii) of reproducibility by a single algorithm. In addition, the necessity of expert judgement means that the mapping of names to clades cannot be automated. This means that the Linnaean system also fails to meet requirement (iii).

148 Second, the lack of reproducibility in the Linnaean system leads, over time, to confusion
149 over the taxon concept to which a name is linked. Through time, different experts often apply the
150 same name in different ways due to different interpretations of the original taxon protologue[3], and
151 consequently, the meaning of this name becomes difficult to track. This problem is further
152 exacerbated by purely nomenclatural issues that notoriously plague taxonomy, such as synonymy,
153 homonymy, misapplication, etc. And even though these can often be reconciled (albeit not always
154 easily) by taxonomic name resolution services (Boyle et al. 2013; Chamberlain and Szöcs 2013),
155 this provides little relief to the long-standing informatics challenge of reconciling names with
156 taxon concepts. This problem is particularly heightened in names with a long history and legacy
157 of taxonomic literature. Because repurposed Linnaean names still point to traditionally
158 circumscribed groups that are not generated in an evolutionary framework, they inherit these
159 problems. In that sense, repurposed Linnaean names approximate to clades, but never exactly
160 match them. This is because traditional groups and the clades we discover are fundamentally two
161 different entities, created by very different criteria (Cellinese et al. 2012). Furthermore, even if the
162 extension of a Linnaean name were to coincide with that of a particular clade, over time this would
163 quickly fall prey to the same problems of interpretation and taxonomic as well as phylogenetic
164 revision. Due to the above points, the Linnaean system fails requirement (i), i.e. it cannot maintain
165 conceptual consistency.
166 Third, the hybrid system still links data to a Linnaean *name*. These names are text strings
167 without computational meaning. Thus, even if we repurpose a Linnaean name to refer to a clade,
168 this name can never express the semantics of that clade. Instead of defining the taxon in a way that
169 would allow machines to identify the taxon, these names link to type specimens and descriptions
170 that, as described above, have been used and interpreted in different ways by different researchers.
171 Thus, as long as Linnaean names are used to point to taxon concepts, it will be impossible for
172 machines to reliably integrate data. This means, again, that the hybrid Linnaean system inevitably
173 fails to meet the requirement of making taxon definitions computationally automatable (iii).
174 The failure of the Linnaean system to meet these three requirements is easiest to explain
175 by drawing an analogy with geolocation-linked data: like taxa, such location data is incredibly

[3] A taxon protologue is the collection of material associated with the publication of a taxon name and concept and therefore, includes all the evidence that support the establishment of a new named entity (e.g., diagnosis, specimens, phylogeny, etc.).

176 useful for integrating data. Imagine that for geolocation-linked data only place names, not standard
177 latitude/longitude geo-coordinates, were available for computation. Data could not be aggregated
178 by region, users could not draw a bounding box on a map to query a database, species occurrence
179 data could not be queried for "all species within 50 miles of my location", and users querying by
180 place would have to know country, state, and possibly city to make the query less ambiguous. Yet,
181 this is the current situation in computing with taxon-linked data.

182 Consider, as an example to illustrate the problems of the Linnaean system, the genus
183 *Campanula* formalized by Linnaeus in 1753, for which *Campanula latifolia* L. was later selected
184 as a lectotype (Britton and Brown 1913). When discussing *Campanula* L., Lammers (2007) states
185 that "there is no modern classification which accounts for this large genus in its entirety" and
186 therefore, the exact number of species is unknown, but the current count is at more than 400. The
187 original description applied to *Campanula* has been so stretched through time that, unsurprisingly,
188 *Campanula* as a Linnaean taxon concept is highly polyphyletic, scattered across the entire
189 Campanuloideae tree with other polyphyletic genera (Crowl et al. 2016; Fig. 2). The clade
190 including the type specimen (*Campanula latifolia*) would have to retain the original name, which
191 would imply a cascade of name changes across the tree, not an uncommon repercussion in
192 taxonomic revisions. Even ignoring the nuisance of name changes, all phylogenetic studies to date
193 have analyzed a significantly incomplete taxon sample, which had stalled any formal update in the
194 taxonomy and classification because it would be premature. The most challenging bottleneck is
195 the inability to retrieve taxonomic concepts unambiguously. Aside from its type specimen, what
196 constitutes the traditional taxon *Campanula,* in view of how the name has been applied across
197 time, is not even easy to verbalize, given an author's subjective taxon description and the lack of
198 informative synapomorphies. Figure 2 illustrates some of the practical consequences of this
199 complex issue, by requesting occurrence data from GBIF (gbif.org) using a query for *Campanula*
200 as *a genus*. Integrating data obtained in this way with the known phylogeny will necessarily be
201 very challenging at best, given that *Campanula* as *a clade* does not exist.

202 Examples like *Campanula* are very common across all domains at any taxonomic level,
203 and the harmonization between traditional ideas about life and the phylogenetic approaches we
204 employ to discover natural entities has become a true impediment to progress in querying,
205 communicating, and 'decorating' all of the parts of the Tree of Life in a consistent and reproducible

206 way. In the next section, we discuss an alternative way of defining taxon concepts for data
207 integration that does not suffer from the problems of the Linnaean system.
208
209 **3. The richness of Phylogenetic Definitions**
210 Starting in the mid 1980's a number of authors suggested that taxon names could be defined
211 by reference to a part of a phylogenetic tree, prompting an extensive theoretical discussion, as well
212 as the first attempts to generate phylogenetic definitions (Ghiselin 1984; Gauthier and Padian
213 1985; Gauthier 1986; Rowe 1987; de Queiroz 1987, 1988; Gauthier et al. 1988; Estes et al. 1988).
214 A phylogenetic definition represents a formal statement that describes a clade in a phylogeny. This
215 body of work laid the foundation for phylogenetic taxonomy, later renamed phylogenetic
216 nomenclature, which takes a strictly tree-thinking approach to biological nomenclature (de
217 Queiroz and Gauthier 1990, 1992, 1994). Soon thereafter, the *PhyloCode* (www.phylocode.org)
218 was drafted as an application of phylogenetic nomenclature's principles.
219 Many systematics papers (e.g., de Queiroz 1992, 1994, 1997; Rowe and Gauthier 1992;
220 Judd et al. 1993, 1994; Bryant 1996, 1997; Sundberg and Pleijel 1994; Christoffersen 1995;
221 Schander and Thollesson 1995; Lee 1996, 1998, 2001; Wyss and Meng 1996; Brochu 1997;
222 Cantino et al. 1997, 2007; Kron 1997; Baum et al. 1998; Eriksson et al. 1998; Härlin and Sundberg
223 1998; Hibbett and Donoghue 1998; Alverson et al. 1999; Pleijel 1999; Sereno 1999; Bremer 2000;
224 Brochu and Sumrall 2001) clearly articulated the need to communicate parts of the Tree of Life
225 and demonstrated that Life could be described by using three basic clade types and their associated
226 phylogenetic definitions. These are (1) minimum clade definitions, denoting the smallest clade that
227 includes the most recent common ancestor, and all its descendants, of two or more internal
228 specifiers; (2) maximum clade definitions, denoting the largest clade that includes the first
229 ancestor, and all its descendants, of one or more internal specifiers but excludes one or more
230 external specifiers; and (3) apomorphy-based definitions, denoting the clade that arises from the
231 first ancestor, and includes all its descendants, that possesses a specified character that is
232 synapomorphic with an internal specifier (Fig. 1). Specifiers are reference points in the phylogeny
233 that serve as anchors for the clade definition and these can be species, specimens, or apomorphies,
234 which would include molecular sequences. Ideally, when using species as specifiers, these would
235 already have a phylogenetic definition available or the Linnaean type present in the phylogeny;

likewise, when using apomorphies, ideally every trait used as specifier should be semantically defined.

While there has been extensive debate in the literature (Benton 2000; Blackwell 2002; Schuh 2003; Polaszek and Wilson 2005; Rieppel 2006; Stevens 2006; de Queiroz and Donoghue 2011; among many others) about possible advantages and disadvantages of the PhyloCode as a nomenclatural system, the PhyloCode is simply one application of phylogenetic nomenclature, in the realm of nomenclatural codes. Our concern here is not arguing the merits of, or issues with the PhyloCode, or, for that matter, any nomenclatural code. Instead, we posit that phylogenetic definitions have unquestionable benefits as a means to unambiguously label all clades in the Tree of Life, and use these for data integration.

Compared to traditional taxon descriptions, phylogenetic definitions have clear advantages for computing with taxon concepts in a phylogenetic context. They draw unambiguous reference to any part of the Tree of Life and can be expressed in a formal and standardized format. Although when published they refer to a taxon concept (clade) originating from a specific phylogenetic topology, a formal clade concept established by an author is an unambiguous statement and approach to communicate taxa, and thus data for those taxa, regardless of future changes in phylogenetic knowledge. That is, as long as the specifiers used in a clade definition have been matched to a given phylogenetic tree, there is no arguing about the clade identified by the definition[4]. Obviously, this cannot prevent or resolve disagreements about the actual taxon concept, but it does enable clearly articulating which element(s) of a phylogenetic definition is(are) the point(s) of contention. In other words, disagreement over a concept does not imply ambiguity over what the concept represents. Additionally, a change in phylogenetic knowledge after the original publication of a phylogenetically defined clade concept may result in taxa now included in the clade that the original author did not intend to be included, or for which the community is divided about the merits of their inclusion. Definitions constructed in some ways will prove more robust, in the judgement of the community, than those built in other ways. However, whether judged "robust" and agreed upon or not, phylogenetic definitions will always unambiguously point to the same clade on any tree containing all its specifiers. For example, our definition of Campanulaceae is "the clade originating with the most recent common ancestor of *Campanula*

[4] We come back to the problem of matching specifiers in section 6.1.

latifolia Linnaeus and all extant organisms or species that share a more recent common ancestor with *Campanula latifolia* than with *Roussea simplex* (Rousseaceae) J. E. Smith, *Pentaphragma ellipticum* (Pentaphragmataceae) Poulsen, or *Stylidium graminifolium* (Stylidiaceae) Swartz ex Willdenow" (Fig. 3; Cellinese 2020).

Others may disagree with this definition, however, there is no ambiguity about the concept being referred to, and the clade it would identify on a given phylogeny.

Phylogenetic definitions are not only beneficial at higher (above species), but also at shallow (species or below-species) taxonomic levels. For example, reconciling Linnaean names with polyphyletic taxa, which are very common across all domains of life, is clearly non-trivial. Often, clades can be diagnosed by interesting morphological or genetic synapomorphies. Traditional taxon names offer little help in referring to such clades, especially if, as is very common, type specimens are missing from the analyses. For example, Crowl et al. (2015) found that *Campanula erinus*, a widespread taxon in the Mediterranean basin, nested in a clade of narrow Aegean archipelago endemics, is polyphyletic and polyploid. In a more in-depth study, Crowl et al. (2017) discovered cryptic diversity within this species due to hybridization with *C. creutzburgii*, which revealed a hybrid lineage that is morphologically identical to *C. erinus*, but differs by having a different ploidy (8x vs the parental 4x). An apomorphy-based clade definition using the trait octoploidy now allows the semantically unambiguous taxonomic recognition of this otherwise cryptic group (Crowl and Cellinese 2017).

Likewise, in other domains, in particular fungi and bacteria, taxa are often so poorly known that only unnamed "phylotypes" can be identified (e.g., Massana et al. 2000; Kim et al. 2012; Lin et al. 2014; Hibbett 2016). Phylogenetic definitions can address these cases, because specifiers can use any uniquely identifiable object suitable for matching the taxonomic unit represented by nodes in a tree. To illustrate this point, in the above Campanulaceae example, the taxonomic unit identified by having scientific name *Campanula latifolia* could also be identified by molecular sequence(s) (e.g., "GenBank: EF141027"), or, as in Crowl and Cellinese (2017), using a specific herbarium specimen with a globally unique identifier.

This potential 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. These entities are not considered 'species' and a clear mechanism to name them is lacking from all of the formal

nomenclature codes. For data publishing, aggregation, and retrieval systems built around names instead of meaning, data for such entities cannot be recovered, certainly not computationally.

These advantages of phylogenetic definitions are widely acknowledged, and phylogenetic definitions have been applied across multiple biological domains in numerous recent phylogenetic studies, resulting in the publication of many clade names, some of which were subsequently repurposed in other analyses (Borchiellini et al. 2004; Joyce et al. 2004; Cantino et al. 2007; Conrad et al. 2011; Soltis et al. 2011; Adl et al. 2012; Cárdenas et al. 2012; Hill et al. 2013; Mannion et al. 2013; Schoch 2013; Sterli et al. 2013; Torres-Carvajal and Mafla-Endara 2013; Wojciechowski 2013; Clemens et al. 2014; Hundt et al. 2014; Rabi et al. 2014; Sferco et al. 2015; Madzia and Cau 2015; Spatafora et al. 2016; Crowl and Cellinese 2017; Wright et al. 2017; Hibbett et al. 2018; de Queiroz et al. 2020; among numerous others). Arguably, this constitutes ample evidence that generating and using taxon concepts defined by patterns of ancestry constitutes an increasing need by the community, and that there is a growing consensus on how to define and use names for such concepts.

4. What is a Phyloreference

In the form commonly published by authors, phylogenetic clade definitions, whether following strict rules of a nomenclatural code (such as the PhyloCode) or not, are natural language text expressions. In this form, the ability to compute with the semantics expressed in the text, as requirement (iii) demands, is severely limited. However, unlike definitions in the Linnaean system, it is possible to transform phylogenetic definitions in natural language text into computable representations and thereby make their semantics accessible to machines. We develop a system for such transformations here, and refer to these computable representations as *phyloreferences*. Specifically, a phyloreference is a representation of a phylogenetic definition as a formal, logic expression that makes its semantics explicit and machine-accessible through the use of terms drawn from ontologies. In this way, phyloreferences are an informatics tool for communicating taxon concepts to machines, as opposed to, for example, a stand-in for Linnaean (or other) nomenclature. As an informatics tool, phyloreferences harness the theoretical, as well as applied, results from a wealth of earlier work in phylogenetic nomenclature to enable machines to integrate and navigate organism-linked data by concepts not afforded by Linnaean taxonomies.

Our proposed approach is based on the Web Ontology Language (OWL 2) (W3C OWL Working Group 2012) Description Logic (DL) framework. OWL has been widely adopted across the life sciences for representing domain knowledge in machine-processable form as ontologies (Mungall et al. 2010, 2011, 2012; Vogt 2009; Jensen and Bork 2010; Deans et al. 2011, 2015; Dahdul et al. 2014; Haendel et al. 2014; Thessen et al. 2015; Senderov et al. 2018). In the context of information science, in which our approach is based, an ontology is a representational model of a knowledge domain, specifically the concepts (represented as classes) comprising the domain, and the relationships that hold between them (represented as relationships between class members). Ontologies have revolutionized our ability to compute with the semantics of natural language expressions. For example, by linking terms in free text phenotype descriptions to formal concepts in community ontologies for the relevant knowledge domains, machine reasoners and statistical algorithms can be used to compute quantitative metrics for the semantic similarity of different phenotype descriptions (Pesquita et al. 2009; Washington et al. 2009; Vision et al. 2011; Bauer et al. 20012; Mabee et al. 2012; Manda et al. 2015; Mabee et al. 2018). Enabling machines to understand the semantics of clade definitions for the purposes of computational data integration is a much less complex task. Nevertheless, clades used by researchers to aggregate or communicate data arguably form part of our body of knowledge about the evolution of the tree of life, and it would thus seem prudent to render it as much computable as other life science knowledge domains.

To afford such capabilities to phylogenetic clade definitions, we propose a model of phyloreferences as defined OWL classes[5]. In this model, the semantics of a phyloreference, and thus the clade concept it represents, are declared by a so-called OWL class expression, which essentially gives the necessary and sufficient conditions for class membership. For a class defined in this way, software tools called reasoners can (among other things) infer for any individual that all individuals that fulfill all conditions necessarily must be instances of the class. We then model the topology of a given phylogeny by declaring its nodes as individuals, and asserting relationships

[5] By class we mean a concept in an ontology, and thus an abstract object (in contrast to individuals or instances, which are concrete objects). Unless stated otherwise, in our use classes have intensional rather than extensional definitions, meaning their descriptions state constraints that must be true for an individual object to be a member of the class. The constraints can be stated in natural language, or as a set of logic conditions. In the latter case, a reasoner can infer class membership.. Similarly, we use the term individual in the sense of an individual member of a group. The usage of this term should not be confused with the question of whether taxa are, in a metaphysical sense, classes or individuals. We hold that, depending on the epistemic context, taxa can be construed as both individuals and kinds (see also Brigandt 2009). Hence, the approach we take here is compatible with the view that taxa are, in a metaphysical sense, individuals.

between those that reflect the topological relationships between nodes. This allows a reasoner to infer which nodes in the phylogeny, if any, match a given phyloreference. This class expression-based model also enables other inferences through computational reasoning. For example, aside from inferring class membership of individuals, OWL reasoners can use these to infer which phyloreferences are equivalent, and which are subclasses of another. Where found, such relationships would be implied solely by the semantics of the clade as represented in the OWL class definition, and as such would hold universally. This is in contrast to approaches that attempt to map Linnaean names to clades in a tree by comparing the clade on the tree and the Linnaean taxon concept based on the relationship (inclusion, overlap, etc.) between their respective sets of members (see "Other Efforts" below).

As argued in the large body of work on phylogenetic nomenclature on which we have based our approach, our proposed models for phyloreference expressions represent patterns of shared and divergent descent, as included and excluded lineages. To illustrate this, a phyloreference for the clade Campanuloideae might be expressed in OWL like this (OWL Manchester Syntax (Horridge and Patel-Schneider 2012); properties in italics; for readability, ontologies of constituent terms are omitted, and term labels are used in place of identifiers):

<Campanuloideae> EquivalentTo *includes_TU* some <Campanula_latifolia> and *excludes_TU* some <Lobelia_cardinalis>.

This expression[6] models a maximum clade definition and asserts that the class Campanuloideae is logically equivalent to the set of nodes that include the taxon concept (TU, for Taxonomic Unit) 'Campanula_latifolia', and exclude the taxon concept 'Lobelia_cardinalis', two necessary and sufficient conditions (called property restrictions in OWL). The properties *includes_TU* and *excludes_TU* are drawn from an ontology, specifically, the Phyloreferencing Ontology, an application ontology that we are developing on top of the Comparative Data Analysis Ontology (CDAO) (Prosdocimi et al. 2009) for defining the semantics of clade definition

[6]The token "some" in the phyloreference example is from OWL Manchester Syntax and signifies existential quantification. Existential quantification (as opposed to universal quantification) properly represents the semantics of the clade definition: for a taxon concept to be included, *some* instance of it needs to be included, not every possibly existing one (observed or not). Likewise for exclusion. TU here is the class of entities that are instances of a given taxon concept. <Campanula_ latifolia> refers to the TU class, "some <Campanula_latifolia>" is some instance of that class.

components. For example, includes_TU as a property is defined such that in the above definition "includes_TU some <Campanula_latifolia>" is true for all nodes that represent an instance of the taxon concept *Campanula latifolia*, or from which such a node descends. In contrast, in the above definition "excludes_TU some <Lobelia_cardinalis>" is true for nodes that have a sibling node representing an instance of the taxon concept *Lobelia cardinalis*, or from which such a node descends. The semantics of a definition with these properties are transparent, unambiguous, and readable by machines. As an ontology class, the definition does not pinpoint one particular node in one particular taxonomy or phylogeny, but the set of all nodes that satisfy the definition. Because the definition is a formal logic expression, class membership can be inferred computationally by a reasoner.

By defining phyloreferences as ontology classes, their adoption, reuse, unambiguous reference, and even community vetting can be promoted using the same mechanisms as for other widely used community ontologies in the life sciences. Specifically, they can be given a label, allowing reference to them by name; assigned globally unique identifiers, making them unambiguously referenceable; and assembled into an ontology maintained in an infrastructure, such as a Github repository that facilitates version control, releases, and community collaboration.

Ultimately, a phyloreference in our approach bears the following important properties. Foremost, it meets our four requirements. Its semantics are unambiguous and machine interpretable because they are expressed in formal logics with uniquely identified ontology terms. This enables reproducing their mapping to a given phylogeny with a fully computational algorithm (requirements (ii) and (iii), and enables maintaining semantic consistency when mapped to different (such as updated) phylogenies (requirement (i)). When a phyloreference is applied to a particular phylogeny that lacks a clade with consistent semantics, there will not be a node that "matches" (i.e. can be inferred as an instance). As a logically defined ontology class, a phyloreference can but need not be named. If it is named, the name is only a label to aid human communication, and this label does not carry semantics a machine is expected to recognize. Phyloreferencing can thus be applied to any branch of the Tree of Life, whether useful names exist or not (requirement (iv)). A phyloreference class can be given a globally unique identifier by which to unambiguously reference it for machines, independent of whether it has a label.

Furthermore, in this way phyloreferences are quite similar to terms in other community ontologies, and our system therefore interoperates naturally with the communities of practice and

409 tool ecosystems that have developed around collections of ontologies in different domains, in
410 particular in the life sciences (Smith et al. 2007).
411
412 **5. Other Efforts to improve the computability of taxon concepts**
413 Even though there has been much controversy over the application of phylogenetic
414 nomenclature (Benton 2000; Blackwell 2002; Schuh 2003; Polaszek and Wilson 2005; Rieppel
415 2006; Stevens 2006; de Queiroz and Donoghue 2013; among many others), its potential to define
416 taxon concept semantics in a logical manner with unambiguously expressible meaning has been
417 recognized before. Hibbet et al. (2005), Keesey (2007), and in part Sereno (2005) and Sereno et
418 al. (2005), already envisioned mechanisms and applications that would leverage computable clade
419 definitions to unambiguously retrieve taxa based on shared descent-based specifications. Keesey
420 (2007) includes a notation and formalism for defining clade names based on mathematical set
421 theory and operators, using the Mathematical Markup Language (MathML), an XML derivative,
422 and extensions to it. Keesey's approach, unlike ours, also supports group concepts that are not
423 monophyletic. However, because MathML is a structured syntax language, not a formal logic,
424 Keesey's approach requires defining custom, bespoke semantics for his notations. It also does not
425 lend itself to publishing clade definitions in the form of ontologies that are readily interoperable
426 with the wealth of other community ontologies increasingly widely used in biology, and the
427 software support even for only reading and interpreting MathML is limited. In practice, Keesey's
428 proposal has not been adopted.
429 Thau and Ludäscher (2007) and Thau et al. (2008) proposed to use Region Connection
430 Calculus (RCC, specifically RCC-5; Randell et al 1992) as a formal logic for computationally
431 reconciling different Linnaean taxonomies (or taxonomic checklists derived from such
432 taxonomies) with each other. RCC-5 defines five basic relationships between two entities:
433 equality, proper inclusion, inverse proper inclusion, overlap, and disjointness. In their approach,
434 human experts assert which relationship(s), called articulations, hold between the concepts from
435 different input taxonomies, such as concepts with identical names, or names that exist in only some
436 of the input taxonomies. Experts also assign or relax a number of so-called global (or latent)
437 taxonomic constraints, such as disjointness of sibling taxa, and parent taxon coverage (every
438 member of a parent taxon is a member of some child taxon). Thau et al. (2008) show that certain
439 machine reasoners can prove the consistency (or inconsistency) of different taxonomies under the

asserted articulations and constraints, and can infer minimally informative relationships (a disjunction of one or more of the RCC-5 base relationships) between concepts.

More recently, Franz et al. (2016, 2019) and Cheng et al. (2017) applied this approach to a variety of complex biological use cases, and also extended it to the challenge of reconciling concepts from traditional Linnaean nomenclature with clades in a phylogenetic tree, as well as aligning clade concepts from competing phylogenetic hypotheses. Although evidently useful for the problem of computationally reconciling taxon concepts, for each new input taxonomy or phylogenetic hypothesis to be reconciled, a considerable amount of effort from trained human experts is necessary to create the articulations and constraints, and the resulting assertions still do not disambiguate or make computable the original intensional semantics of a taxon concept. Therefore, it does not make the exercise of repurposing Linnaean names for clades in a phylogenetic tree a less subjective and manual approximation than it necessarily is, because the concepts at hand are fundamentally different in nature.

6. Challenges and Limitations

Previous proposals to replace the Linnaean system with a purely phylogenetic alternative have proven to be very controversial. As our proposal does not concern taxonomic nomenclature or classification, many of these controversies are not directly relevant. However, there are various ways in which opponents might object against the arguments in this paper. We respond to these briefly, and point to limitations and challenges for our approach.

6.1. Specifiers

One of the greater challenges in applying phyloreferences on a larger scale, and across different phylogenetic trees, is that phylogenetic clade definitions are "anchored" by the specifiers designating the taxon concepts that are to be included or excluded. Therefore, resolving a phyloreference on a tree necessarily requires that the anchoring taxon concepts of a phyloreference, and the taxon concepts linked to (typically terminal) nodes in a phylogeny, can be "matched" by a reasoner. More specifically, these taxon concepts need to be defined such that the reasoner can infer when a taxon concept used in the phyloreference is congruent with, or includes, a taxon concept linked to a tree node. In some cases such a match will be exact and unambiguous, for example, if the specifier and node-linked taxon concept are referenced to the same globally

unique identifier. In practice, matching specifiers between phyloreference and phylogeny is an inherently non-trivial problem, and matches will range from unambiguous to approximate. For example, if taxon concept references are, as will commonly be the case, Linnaean taxon names, even an exact match is not necessarily free of ambiguity, such as when the names are not demonstrably drawn from the same taxonomy. Indeed, this is the taxonomic name resolution problem that arises whenever Linnaean taxon names must be reconciled, and the confidence in name matches will follow the familiar spectrum. Especially for phylogenies with incomplete taxon sampling, a taxon concept used as specifier in a phyloreference may also be altogether absent from a tree. The question is, then, whether or not one of the taxon concepts present on the tree can substitute for the specifier without changing the semantics of the clade definition. Whether this is possible or not will in turn depend on the definition of the clade and the phylogeny at hand on which it is to be recovered, and may require sophisticated algorithms to determine.

Phyloreferences by themselves do not obviate the need to match or reconcile Linnaean taxon names. However, this is due to the prevailing practice of identifying taxon concepts through names, rather than a specific weakness in the phyloreferencing approach; and because phyloreferences are in essence uniquely identifiable ontology terms, this problem and the ambiguity it confers are not re-introduced every time data are linked to a taxon. Furthermore, how and why a taxon concept for a specifier matches one for a node in a tree can be expressed through formal axioms in the same logic framework (i.e., OWL2 in our case), and thus be documented in a fully reproducible manner. For example, if a target phylogeny lacks a node for *Campanula latifolia*, but contains a node for *Campanula*, a "mapping" axiom asserting that the concept *Campanula* includes *Campanula latifolia* will allow matching a phyloreference for the Campanuloideae clade that references *Campanula latifolia* as a specifier that must be included.

Finally, it is worth emphasizing that the ambiguity inherent in reconciling names by itself does not introduce ambiguity into the semantics of the clade definition, though it does render *recovering* the clade semantics on phylogenies, other than the one used by the original author, prone to the same problems that beset taxon name matching in general. Creating mapping axioms in an effective and scalable manner may be non-trivial, but we are confident that solutions to address this challenge can and will be developed. In the meantime, the Open Tree of Life offers a comprehensive, even if synthetic, phylogeny that is continuously updated with evolving

phylogenetic knowledge, and with names for terminal nodes sourced from dozens of taxonomies (Rees and Cranston 2017).

6.2. Genealogical discordance

It is well-known that, due to phenomena such as lateral gene transfer, hybridization, introgression, and others, evolution is often not tree-like across all domains of life, including Archaea, bacteria and fungi. One might worry then that the phyloreferences proposed here are not suitable for capturing groups whose evolutionary relations are more suitably represented by a network than by a bifurcating pattern. Although phylogenies are hierarchical, with clades that are either nested or mutually exclusive, reticulation due to different biological processes results in partially overlapping clades, with hybrid lineages belonging to both parental clades. Partially overlapping clades can, in fact, be phylogenetically defined, which demonstrates the flexibility of this approach. For example, Crowl and Cellinese (2017) illustrate how phylogenetic definitions apply to lineages derived from hybridization and polyploidy (using ploidy in an apomorphy-based definition), and allow the naming of cryptic diversity.

Phylogenetic reconstructions may generate discordant hypotheses that are best synthesized by networks rather than bifurcating patterns. For considering the question whether phyloreferences can be meaningfully applied to such networks, note that in principle the key concepts used in our approach for encoding the semantics of a clade definition, namely ancestors and descendants, and taxon concepts included in or excluded from a line of descendents, still fully apply in networks. Hence, there is no theoretical or technical reason that would prevent resolving a phyloreference on a phylogenetic network. Nonetheless, a clade retrieved in this way should be treated with great caution, because at least for now the underlying clade definition will have almost universally been erected based on a phylogenetic tree, not a network. Therefore, the benefit of applying phyloreferences to networks as part of, for example, a data integration project, seems questionable at best.

6.3. Adoption cost

One could object that even if phyloreferences are in principle preferable over Linnaean names for integrating data, the cost of adoption would be very high, or high enough to outweigh the benefits. For a response, we note but set aside the fact that such an argument would attribute

limited value to the problems caused by using the Linnaean system; we disagree that irreproducible science has only limited costs. Nonetheless, we acknowledge that as for any novel system for indexing data, for a resource such as GBIF, with huge amounts of data that need to be queryable very efficiently by a large user community, to fully support phyloreferencing would likely have a significant engineering cost. This notwithstanding, we find it important to note that phyloreferences can already be taken advantage of right now, including for data integration projects, by tapping into and combining already existing technologies. To sketch out an example, the programming interface (API) to the Open Tree of Life includes a most recent common ancestor query service that depending on the input parameters returns the common ancestor node semantically fully consistent with minimum clade and maximum clade definitions, respectively, that underlie phyloreferences. Additional Open Tree of Life query services can then be used to obtain the species contained by the clade resolved in the previous step, which then in turn allow querying a database indexed by Linnaean names for data associated with the clade. This approach can already be used, for example, to find how phylogenetic clades vs Linnaean names can result in different inferences, such as geographical distribution.

7. Final remarks

We strongly believe we are at a crossroad where the idiosyncratic applications of Linnaean nomenclature and taxonomy to the approach we use to discover and name taxa is simply untenable in the age of computationally-driven science. Linnaean names represent an incurable theoretical and practical shortfall. We suggest that phyloreferencing lays the foundation for an informatics infrastructure that enables using the Tree of Life to organize, query, and navigate our knowledge of biodiversity. Building this foundation now is timely. Large phylogenies encompassing diverse groups across the tree of life are published in increasing numbers (e.g., Smith et al. 2011; Hinchliff et al. 2015; Smith and Brown 2018; Howard et al. 2019). Especially for large tree synthesis projects, the need for phyloreferencing has already arisen, because it is the basis for persistently and reproducibly linking data and metadata to internal nodes (i.e. clades) in the tree. There are also parts of the Tree of Life for which a stunning organismal and trait diversity is only just beginning to be characterized, and for which the traditional fallback of Linnaean names is hardly available, and perhaps never will be (e.g., microbial diversity, and population-level diversity). 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 and concepts that allow researchers to communicate data and knowledge about their groups, yet also have fully computable and thus reproducible semantics built-in.

One of the key goals of phyloreferences is to enable computationally querying, navigating, integrating, and visualizing any data linked to groups of organisms, in a way that is driven by evolutionary relatedness. We have argued that merely repurposing Linnaean names onto trees cannot achieve this goal. Phyloreferences allow us to compare parts of the Tree of Life about which we would otherwise not be able to communicate. Consequently, the number of phylogenetic taxon definitions being published has already increased rapidly in recent years across multiple domains, signifying that phylogenetic approaches to diagnose taxonomic groups and their names are being increasingly widely adopted and ideally, every clade discovered should bear a definition. When translated into formal phyloreferences, the semantics of these definitions not only become fully accessible to machines, but by curating them into a community ontology, they become much more findable and reusable compared to when buried in the text of publications.

We believe that a phylogenetic data synthesis encompasses far more than a challenging topological synthesis. The approach we propose is native to tree-thinking and completely flexible because phyloreferences adapt seamlessly to changes in phylogenetic knowledge and would therefore apply to small and large topologies and syntheses. In view of the upcoming publication of the PhyloCode and the ever-increasing number of published phylogenetic definitions, now is the time to envision the Tree of Life as a navigable map where clade definitions (taxon concepts) serve as physical addresses and phyloreferences provide the means to achieve a retraceable navigation.

References

Adl S.M., Simpson A.G.B., Lane C.E., Lukeš J., Bass D., Bowser S.S., Brown M.W., Burki F., Dunthorn M., Hampl V., Heiss A., Hoppenrath M., Lara E., Gall L.L., Lynn D.H., McManus H., Mitchell E.A.D., Mozley-Stanridge S.E., Parfrey L.W., Pawlowski J.,Rueckert S., Shadwick L., Schoch C.L., Smirnov A., Spiegel F.W. 2012. The Revised Classification of Eukaryotes. J. Eukaryot. Microbiol. 59:429–493.

Allen J.M., Folk R.A., Soltis P.S., Soltis D.E., Guralnick R.P. 2018. Biodiversity synthesis across the green branches of the tree of life. Nat. Plants 5:11-13.

Alverson W.S., Whitlock B.A., Nyffeler R., Bayer C., Baum D.A. 1999. Phylogeny of the core Malvales: evidence from ndhF sequence data. Am. J. Bot. 86:1474–1486.

Bauer S., Köhler S., Schulz M.H., Robinson P.N. 2012. Bayesian ontology querying for accurate and noise-tolerant semantic searches. Bioinformatics 28:2502–2508.

Baum D.A., Alverson W.S., Nyffeler R. 1998. A durian by any other name: taxonomy and nomenclature of the core Malvales. Harv. Pap. Bot. 3:315–330.

Beaman R.S., Cellinese N. 2012. Mass digitization of scientific collections: New opportunities to transform the use of biological specimens and underwrite biodiversity science. ZooKeys 209:7–17.

Benton M.J. 2000. Stems, nodes, crown clades, and rank-free lists: is Linnaeus dead? Biol. Rev. 75:633–648.

Blackwell J.H. 2002. One-hundred-year code déjà vu? Taxon 51:151–154.

Borchiellini C., Chombard C., Manuel M., Alivon E., Vacelet J., Boury-Esnault N. 2004. Molecular phylogeny of Demospongiae: implications for classification and scenarios of character evolution. Mol. Phylogenet. Evol. 32:823–837.

Boyle B., Hopkins N., Lu Z., Garay J.A.R., Mozzherin D., Rees T., Matasci N., Narro M.L., Piel W.H., Mckay S.J., Lowry S., Freeland C., Peet R.K., Enquist B.J. 2013. The taxonomic name resolution service: an online tool for automated standardization of plant names. BMC Bioinformatics. 14:16.

Bremer K. 2000. Phylogenetic nomenclature and the new ordinal system of the angiosperms. In: Nordenstam B., El Ghazaly G., Kassas M., editors. Plant systematics for the 21st century. London, UK: Portland Press. p. 125–133.

624 Brigandt I. 2009. Natural kinds in evolution and systematics: Metaphysical and epistemological
625 considerations. Acta Biotheoretica, 57(1–2): 77–97.
626 Britton C.E., Brown A. 1913. An illustrated flora of the northern United States. Edition 2, Volume
627 3, p. 294.
628 Brochu CA. 1997. Synonymy, redundancy, and the name of the crocodile stem-group. J. Vertebr.
629 Paleontol. 17:448–449.
630 Brochu C.A., Sumrall C.D. 2001. Phylogenetic nomenclature and paleontology. J. Paleontol.
631 75:754–757.
632 Brook B.W., Sodhi N.S., Bradshaw C.J.A. 2008. Syndergies among extinction drivers under global
633 change. Trends Ecol. Evol. 23:453–460.
634 Brown J.H., Lomolino M.V. 1998. Biogeography. Sunderland, MA: Sinauer.
635 Bryant H.N. 1996. Explicitness, Stability, and Universality in the Phylogenetic Definition and
636 Usage of Taxon Names: A Case Study of the Phylogenetic Taxonomy of the Carnivora
637 (Mammalia). Syst. Biol. 45:174–189.
638 Bryant H.N. 1997. Cladistic information in phylogenetic definitions and designated phylogenetic
639 contexts for the use of taxon names. Biol. J. Linn. Soc. 62:495–503.
640 Cantino P.D., Olmstead R.G., Wagstaff S.J. 1997. A Comparison of Phylogenetic Nomenclature
641 with the Current System: A Botanical Case Study. Syst. Biol. 46:313–331.
642 Cantino P.D., Doyle J.A., Graham S.W., Judd W.S., Olmstead R.G., Soltis D.E., Soltis P.S.,
643 Donoghue M.J. 2007. Towards a phylogenetic nomenclature of Tracheophyta. Taxon 56:822-
644 846.
645 Cárdenas P., Pérez T., Boury-Esnault N. 2012. Sponge systematics facing new challenges. In:
646 Becerro M.A., Uriz M.J., Maldonado M., Turon X., editors. Advances in Sponge Science:
647 Phylogeny, Systematics, Ecology. London, UK: Academic Press. p. 79–209.
648 Cellinese N., Baum D.A., Mishler B.D. 2012. Species and Phylogenetic Nomenclature. Syst. Biol.
649 61:885–891.
650 Cellinese N. 2020. Campanulaceae. In de Queiroz K., Cantino P.D., Gauthier J., editors.
651 Phylonyms: a Companion to the PhyloCode, pp. 381-383. CRC Press, Boca Raton, FL.
652 Chamberlain S.A., Szöcs E. 2013. Taxize: taxonomic search and retrieval in R. F10000Res 2:191.
653 Chan C.X., Ragan M.A. 2013. Next-generation phylogenomics. Biology Direct 8:3.

654 Cheng Y.Y., Franz N., Schneider J., Yu S., Rodenhausen T., Ludäscher B. 2017. Agreeing to
655 disagree: Reconciling conflicting taxonomic views using a logic-based approach. Proc. Assoc.
656 Inf. Sci. Technol. 54:46–56.
657 Christoffersen M.L. 1995. Cladistic Taxonomy, Phylogenetic Systematics, and Evolutionary
658 Ranking. Syst. Biol. 44:440–454.
659 Clemens W.L., Arakaki M., Sweeney P.W., Edwards E.J., Donoghue, M.J. 2014. A chloroplast
660 tree for *Viburnum* (Adoxaceae)) and its implications for phylogenetic classification and
661 character evolution. Am. J. Bot. 101:1029–1049.
662 Conrad J.L., Ast J.C., Montanari S., Norell M.A. 2011. A combined evidence phylogenetic
663 analysis of Anguimorpha (Reptilia: Squamata). Cladistics 27:230–277.
664 Crowl A.A.**,** Visger C., Mansion G., Hand R., Wu H.-H., Kamari G., Phitos D., Cellinese N. 2015.
665 Evolution and Biogeography of the Endemic *Roucela* complex (Campanulaceae: Campanula)
666 in the Eastern Mediterranean. Ecol. Evol. 10.1002/ece3.179.
667 Crowl A.A., Miles N.**,** Visger C., Hansen K., Ayers T., Haberle R., Cellinese N. 2016. A global
668 perspective on Campanulaceae: biogeographic, genomic, and floral evolution. Am. J.
669 Bot. 103:233–245.
670 Crowl A.A., Myers C., Cellinese N. 2017. Embracing discordance: Phylogenomic analyses
671 provide evidence for allopolyploidy leading to cryptic diversity in a Mediterranean *Campanula*
672 (Campanulaceae) clade. Evolution 71:913–922.
673 Crowl A.A., Cellinese N. 2017. Naming diversity in an evolutionary context: Phylogenetic
674 definitions of the *Roucela* clade (Campanulaceae/Campanuloideae) and the cryptic taxa within.
675 Ecol. Evol. 10.1002/ece3.3442.
676 Dahdul W.M., Cui H., Mabee P.M., Mungall C.J., Osumi-Sutherland D., Walls R.L., Haendel
677 M.A. 2014. Nose to tail, roots to shoots: spatial descriptors for phenotypic diversity in the
678 Biological Spatial Ontology. J. Biomed. Semantics 5:34.
679 Darwin C. 1859. On the Origin of Species. London (UK): John Murray.
680 de Queiroz K. 1987. Phylogenetic systematics of iguanine lizards. A comparative osteological
681 study. Univ. Calif. Publ. Zool. 118:1–203.
682 de Queiroz K. 1988. Systematics and the Darwinian Revolution. Philos. Sci. 55:238–259.
683 de Queiroz K., Gauthier J.A. 1990. Phylogeny as a Central Principle in Taxonomy: Phylogenetic
684 Definitions of Taxon Names. Syst. Zool. 39:307–322.

685 de Queiroz K. 1992. Phylogenetic definitions and taxonomic philosophy. Biol. Philos. 7:295–313.
686 de Queiroz K., Gauthier J.A. 1992. Phylogenetic Taxonomy. Annu. Rev. Ecol. Syst. 23:449–480.
687 de Queiroz K. 1994. Replacement of an Essentialistic Perspective on Taxonomic Definitions as
688 Exemplified by the Definition of "Mammalia". Syst. Biol. 43:497–510.
689 de Queiroz K., Gauthier J.A. 1994. Toward a phylogenetic system of biological nomenclature.
690 Trends Ecol. Evol. 9:27–31.
691 de Queiroz K. 1997. The Linnaean hierarchy and the evolutionization of taxonomy, with emphasis
692 on the problem of nomenclature. Aliso 15:125–144.
693 de Queiroz K., Donoghue M.J. 2011. Phylogenetic Nomenclature, Three-Taxon Statements, and
694 Unnecessary Name Changes. Syst. Biol. 60:887–892.
695 de Queiroz K., Donoghue M.J. 2013. Phylogenetic Nomenclature, Hirerachical Information, and
696 Testability. Syst. Biol. 62:167–174.
697 De Queiroz K., Cantino P.D., Gauthier J.A. 2020. Phylonyms: a companion to the PhyloCode.
698 CRC Press, Boca Raton, Florida., FL.
699 Deans A.R., Yoder M.J., Balhoff J.P. 2011. Time to change how we describe biodiversity. Trends
700 Ecol. Evol. 27:78–84.
701 Deans A.R., Lewis S.E., Huala E., Anzaldo S.S., Ashburner M., Balhoff J.P., Blackburn D.C.,
702 Blake J.A., Burleigh J.G., Chanet B., Cooper L.D., Courtot M., Csösz S., Cui H., Dahdul W.,
703 Das S., Dececchi T.A., Dettai A., Diogo R., Druzinsky R.E., Dumontier M., Franz N.M.,
704 Friedrich F., Gkoutos G.V., Haendel M., Harmon L.J., Hayamizu T.F., He Y., Hines H.M.,
705 Ibrahim N., Jackson L.M., Jaiswal P., James-Zorn C., Köhler S., Lecointre G., Lapp H.,
706 Lawrence C.J., Le Novère N., Lundberg J.G., Macklin J., Mast A.R., Midford P.E., Mikó I.,
707 Mungall C.J., Oellrich A., Osumi-Sutherland D., Parkinson H., Ramírez M.J., Richter S.,
708 Robinson P.N., Ruttenberg A., Schulz K.S., Segerdell E., Seltmann K.C., Sharkey M.J., Smith
709 A.D., Smith B., Specht C.D., Squires R.B., Thacker R.W., Thessen A., Fernandez-Triana J.,
710 Vihinen M., Vize P.D., Vogt L., Wall C.E., Walls R.L., Westerfeld M., Wharton R.A., Wirkner
711 C.S., Woolley J.B., Yoder M.J., Zorn A.M., Mabee P. 2015. Finding Our Way through
712 Phenotypes. PLoS Biol. 13:e1002033
713 Eliason C.M., Andersen M.J., Hackett S.J. 2019. Using Historical Biogeography Models to Study
714 Color Pattern Evolution. Syst. Biol. syz012 doi.org/10.1093/sysbio/syz012

715 Ereshefsky M. 2001. The Poverty of the Linnaean Hierarchy. A phylosohical study of biological
716 taxonomy. Cambridge (MA): Cambridge University Press.
717 Eriksson T., Donoghue M.J., Hibbs M.S. 1998. Phylogenetic analysis of *Potentilla* using DNA
718 sequences of nuclear ribosomal internal transcribed spacers (ITS), and implications for the
719 classification of Rosoideae (Rosaceae). Plant Syst. Evol. 211:155–179.
720 Estes R., de Queiroz K., Gauthier J. 1988. Phylogenetic relationships within Squamata. In: Estes
721 R., Pregill G.K., editors. Phylogenetic relationships of the lizard families: essays
722 commemorating Charles L. Camp. Stanford (CA): Stanford University Press. p. 119–281.
723 Folk R.A., Stubbs R.L., Mort M.E., Cellinese N., Allen J.M., Soltis P.S., Soltis D.E., Guralnick
724 R.P. 2019. Rates of niche and phenotype evolution lag behind diversification in a temperate
725 radiation. Proc. Nat. Acad. Sci. 116:10874–10882.
726 Franz N.M., Chen M., Kianmajd P., Yu S., Bowers S., Weakley A.S., Ludäscher B. 2016. Names
727 are not good enough: reasoning over taxonomic change in the *Andropogon* complex 1. Semant.
728 Web 7:645–667.
729 Franz N.M., Musher L.J., Brown J.W., Yu S., Ludäscher B. 2019. Verbalizing phylogenomic
730 conflict: Representation of node congruence across competing reconstructions of the neoavian
731 explosion. PLoS Comput. Biol. 15:e1006493.
732 Gauthier J., Padian K. 1985. Phylogenetic, functional, and aerodynamic analyses of the origin of
733 birds and their flight. In: Hecht K., Ostrom G.H., Viohl G., Wellnhofer P., editors. The
734 beginnings of birds. Eichstatt (Germany): Freude des Jura-Museums. p. 185–197.
735 Gauthier J. 1986. Saurischian monophyly and the origin of birds. In: Padian K., editor. The origin
736 of birds and the evolution of flight. San Francisco (CA): California Academy of Sciences. p.
737 1–55.
738 Gauthier J., Estes R., de Queiroz K. 1988. A phylogenetic analysis of Lepidosauromorpha. In:
739 Estes R., Pregill G.K., editors. Phylogenetic relationships of the lizard families: essays
740 commemorating Charles L. Camp. Stanford (CaA): Stanford University Press. p. 15–98.
741 Ghiselin MT. 1984. "Definition," "Character," and Other Equivocal Terms. Syst. Zool. 33:104–
742 110.
743 Haendel M., Balhoff J., Bastian F., Blackburn D., Blake J., Bradford Y., Comte A., Dahdul W.,
744 Dececchi T., Druzinsky R., Hayamizu T., Ibrahim N., Lewis S., Mabee P., Niknejad A.,

745 Robinson-Rechavi M., Sereno P., Mungall C. 2014. Unification of multi-species vertebrate
746 anatomy ontologies for comparative biology in Uberon. J. Biomed. Semantics 5:21.
747 Hampton S.E., Jones M.B., Wasser L.W., Schildhauer M.P., Supp S.R., Brun J., Hernandez R.R.,
748 Boettiger C., Collins S.L., Gross L.J., Fernández D.S., Budden A., White E.P., Teal T.K.,
749 Labou S.G., Aukema J.E. 2017. Skills and Knowledge for Data-Intensive Environmental
750 Research. *BioScience* 67:546–557.
751 Härlin M., Sundberg P. 1998. Taxonomy and Philosophy of Names. Biol. Philos. 13:233–244.
752 Hennig, W. 1950. Grundzüge einer Theorie der phylogenetischen Systematik. Berlin (Germany):
753 Deutscher Zentralverlag.
754 Hennig W. 1966. Phylogenetic Systematics. Urbana (IL): University of Illinois Press.
755 Hibbett D.S., Donoghue M.J. 1998. Integrating Phylogenetic Analysis and Classification in Fungi.
756 Mycologia 90:347–356.
757 Hibbett D.S., Nilsson R.H., Snyder M., Fonseca M., Costanzo J., Shonfeld M. 2005. Automated
758 Phylogenetic Taxonomy: An Example in the Homobasidiomycetes (Mushroom-Forming
759 Fungi). *Syst. Biol.* 54: 660–668.
760 Hibbett D. 2016. The invisible dimension of fungal diversity. Science 351:1150-1151.
761 Hibbett D.S., Blackwell M., James T.Y., Spatafora J.W., Taylor J.W., Vilgalys R. 2018.
762 Phylogenetic taxon definitions for Fungi, Dikarya, Ascomycota and Basidiomycota. *IMA*
763 *FUNGUS* 9:291–298.
764 Hill M.S., Hill A.L., Lopez J., Peterson K.J., Pomponi S., Diaz M.C., Thacker R.W., Adamska M.,
765 Boury-Esnault N., Cárdenas P., Chaves-Fonnegra A., Danka E., De Laine B.-O., Formica D.,
766 Hajdu E., Lobo-Hajdu G., Klontz S., Morrow C.C., Patel J., Picton B., Pisani D., Pohlmann
767 D., Redmond N.E., Reed J., Richey S., Riesgo A., Rubin E., Russell Z., Rützler K., Sperling
768 E.A., di Stefano M., Tarver J.E., Collins A.G. 2013. Reconstruction of Family-Level
769 Phylogenetic Relationships within Demospongiae (Porifera) Using Nuclear Encoded
770 Housekeeping Genes. PLoS ONE 8:e50437.
771 Hinchliff C.E., Smith S.A., Allman J.F., Burleigh J.G., Chaudhary R., Coghill L.M., Crandall
772 K.A., Deng J., Drew B.T., Gazis R., Gude K., Hibbett D.S., Katz L.A., Laughinghouse
773 H.D., McTavish E.J., Midford P.E., Owen C.L., Ree R.H., Rees J.A., Soltis D.E., Williams
774 T., Cranston K.A. 2015. Synthesis of phylogeny and taxonomy into a comprehensive tree of
775 life. Proc. Nat. Acad. Sci. 112:12764–12769.

776 Horridge M., Patel-Schneider P.F. 2012. OWL 2 Web Ontology Language Manchester Syntax
777 (Second Edition).
778 Hortal J., de Bello F., Diniz-Filho J.A.F., Lewinsohn T.M., Lobo J.M., Ladle R.J. 2015. Seven
779 Shortfalls that Beset Large-Scale Knowledge of Biodiversity. Annu. Rev. Ecol. Evol. Syst.
780 46:523–549.
781 Howard C.C., Folk R., Beaulieu J.M., Cellinese N. 2019. The monocotyledonous underground:
782 global climatic and phylogenetic patterns of geophyte diversity. Am. J. Bot. 106:850–863.
783 Howard C.C., Landis, J.B., Beaulieu J.M., Cellinese N. 2020. Geophytism in monocots lead to
784 higher rates of diversification. New Phytologist 225: 1023-1032.
785 Hundt P.J., Iglésias S.P., Hoey A.S., Simons A.M. 2014. A multilocus molecular phylogeny of
786 combtooth blennies (Percomorpha: Blennioidei: Blenniidae): Multiple invasions of intertidal
787 habitats. Mol. Phylogenet. Evol. 70:47–56.
788 Jensen L.J., Bork P. 2010. Ontologies in quantitative biology: a basis for comparison, integration,
789 and discovery. PLoS Biol. 8:e1000374.
790 Joyce W.G., Parham J.F., Gauthier J.A. 2004. Developing a protocol for the conversion of rank-
791 based taxon names to phylogenetically defined clade names, as exemplified by turtles. J.
792 Paleontol. 78:989–1013.
793 Judd W.S., Stern W., Cheadle V.I. 1993. Phylogenetic position of *Apostasia* and *Neuwiedia*
794 (Orchidaceae). Bot. J. Linn. Soc. 113:87–94.
795 Judd W.S., Sanders R.W., Donoghue M.J. 1994. Angiosperm family pairs: preliminary
796 phylogenetic analyses. Harv. Pap. Bot. 5:1–51.
797 Keesey T.M. 2007. A mathematical approach to defining clade names, with potential applications
798 to computer storage and processing. Zool. Scr. 36:607–621.
799 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
800 H., Won S., Chun J. 2012. Introducing EzTaxon-e: a prokaryotic 16S rRNA gene sequence
801 database with phylotypes that represent uncultured species. Int. J. Syst. Evol. Microbiol.
802 62:716–721.
803 Kozlov A.M., Darriba D., Flouri T., Morel B., Stamatakis A. 2019. RAxML-NG: A fast, scalable,
804 and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics
805 btz305 doi.org/10.1093/bioinformatics/btz305.
806 Kron K.A. 1997. Exploring alternative systems of classification. Aliso 15:105–111.

807 Lammers T.G. 2007. Campanulaceae. In: Kadereit J.W., Jeffrey C., editors. The Families and
808 Genera of Vascular Plants, vol 8. Berlin, Heidelberg: Springer Verlag. p. 26–56.
809 Lee M.S.Y. 1996. Stability in Meaning and Content of Taxon Names: An Evaluation of Crown-
810 Clade Definitions. Proc. R. Soc. Lond. Ser. B. 263:1103–1109.
811 Lee M.S.Y. 1998. Phylogenetic Uncertainty, Molecular Sequences, and the Definition of Taxon
812 Names. Syst. Biol. 47:719–726.
813 Lee M.S.Y. 2001. On Recent Arguments for Phylogenetic Nomenclature. Taxon 50:175–180.
814 Lemmon E.M., Lemmon A.R. 2013. High-Throughput Genomic Data in Systematics and
815 Phylogenetics. Annu. Rev. Ecol. Evol. Syst. 44:99–121.
816 Lin C.H., Tsai K.C., Prior P., Wang J.F. 2014. Phylogenetic relationships and population structure
817 of Ralstonia solanacearum isolated from diverse origins in Taiwan. Plant Pathol. 63:1395–
818 1403.
819 Linnaeus C. 1753. Species Plantarum. Laurentius Salvius, Stockholm.
820 Mabee P., Balhoff J.P., Dahdul W.M., Lapp H., Midford P.E., Vision T.J., Westerfield M. 2012.
821 500,000 fish phenotypes: The new informatics landscape for evolutionary and developmental
822 biology of the vertebrate skeleton. J. Appl. Ichthyol. 28:300–305.
823 Mabee P.M., Dahdul W.M., Balhoff J.P., Lapp H., Manda P., Uyeda J., Vision T., Westerfield M.
824 2018. Phenoscape: Semantic analysis of organismal traits and genes yields insights in
825 evolutionary biology. In: Thessen A.E., editor. Application of Semantic Technology in
826 Biodiversity Science. IOS Press.
827 Madzia D., Cau A. 2017. Inferring ‘weak spots’ in phylogenetic trees: application to mosasauroid
828 nomenclature. PeerJ 5:e3782.
829 Manda P., Balhoff J.P., Lapp H., Mabee P., Vision T.J. 2015. Using the Phenoscape
830 Knowledgebase to relate genetic perturbations to phenotypic evolution. Genesis 53:561–571.
831 Mannion P.D., Upchurch P., Barnes R.N., Mateus O. 2013. Osteology of the Late Jurassic
832 Portuguese sauropod dinosaur *Lusotitan atalaiensis* (Macronaria) and the evolutionary history
833 of basal titanosauriforms. Zool. J. Linn. Soc. 168:98–206.
834 Massana R., DeLong E.F., Pedros-Alio C. 2000. A few cosmopolitan phylotypes dominate
835 planktonic archaeal assemblages in widely different oceanic provinces. Appl. Environ.
836 Microbiol. 66:1777–1787.

837 McTavish E.J., Drew B.T., Redelings B., Cranston K.A. 2017. How and Why to Build a Unified
838 Tree of Life. BioEssays 39: 1700114.
839 Michener W.K., Jones M.B. 2012. Ecoinformatics: supporting ecology as a data-intensive science.
840 Trends Ecol. Evol. 27: 85–93.
841 Mishler, B., & Wilkins, J. S. (2018). The Hunting of the SNaRC: A Snarky Solution to the
842 Species Problem. Philosophy, Theory, and Practice in Biology, 10.
843 Mungall C.J., Gkoutos G.V., Smith C.L., Haendel M.A., Lewis S.E., Ashburner M. 2010.
844 Integrating phenotype ontologies across multiple species. Genome Biol. 11:R2.
845 Mungall C.J., Bada M., Berardini T.Z., Deegan J., Ireland A., Harris M.A., Hill D.P., Lomax J.
846 2011. Cross-product extensions of the Gene Ontology. J. Biomed. Inf. 44:80–86.
847 Mungall C.J., Torniai C., Gkoutos G.V., Lewis S.E., Haendel M.A. 2012. Uberon, an integrative
848 multi-species anatomy ontology. Genome Biol. 13:R5.
849 Page L.M., MacFadden B.J., Fortes J.A., Soltis P.S., Riccardi G. 2015. Digitization of Biodiversity
850 Collections Reveals Biggest Data on Biodiversity. BioScience 65: 841-842.
851 Parr C.S., Guralnick R., Cellinese N., Page R.D.M. 2012. Evolutionary informatics: unifying
852 knowledge about the diversity of life. Trends Ecol. Evol. 27:94–103.
853 Pesquita C., Faria D., Falcão A.O., Lord P, Couto F.M. 2009. Semantic similarity in biomedical
854 ontologies. PLoS Comput. Biol. 5:e1000443.
855 Philippe H., Delsuc F., Brinkmann H., Lartillot N. 2005. Phylogenomics. Annu. Rev. Ecol. Evol.
856 Syst. 36: 541–562.
857 Pleijel F. 1999. Phylogenetic Taxonomy, a Farewell to Species, and a Revision of Heteropodarke
858 (Hesionidae, Polychaeta, Annelida). Syst. Biol. 48:755–789.
859 Polaszek A., Wilson E.O. 2005. Sense and stability in animal names. Trends Ecol. Evol. 20:421–
860 422.
861 Porter J.H., Nagy E., Kratz T.K., Hanson P., Collins S.L., Arzberger P. 2009. New eyes on the
862 world: Advanced sensors for ecology. BioScience 59: 385–397.
863 Prosdocimi F., Chisham B., Pontelli E., Thompson J.D., Stoltzfus A. 2009. Initial Implementation
864 of a comparative Data Analysis Ontology. Evol. Bioinform. Online 47–66.
865 Rabi M., Sukhanov V.B., Egorova V.N., Danilov I., Joyce W.G. 2014. Osteology, relationships,
866 and ecology of *Annemys* (Testudines, Eucryptodira) from the Late Jurassic of Shar Teg,

867 Mongolia, and phylogenetic definitions for Xinjiangchelyidae, Sinemydidae, and
868 Macrobaenidae. J. Vert. Paleontol. 34:327–352.
869 Randell D. A., Cui Z., Cohn A. 1992. A spatial logic based on regions and connection. In: B.
870 Nebel, C. Rich, Swartout W., editors. KR'92. Principles of Knowledge Representation and
871 Reasoning: Proceedings of the Third International Conference. San Mateo (CA): Morgan
872 Kaufmann. P. 165–176.
873 Rees J.A., Cranston K. 2017. Automated assembly of a reference taxonomy for phylogenetic data
874 synthesis. Biodivers Data J. e12581.
875 Rieppel O. 2006. The PhyloCode: a critical discussion of its theoretical foundation. Cladistics
876 22:186–197.
877 Rowe T. 1987. Definition and Diagnosis in the Phylogenetic System. Syst. Zool. 36:208–211.
878 Rowe T., Gauthier J. 1992. Ancestry, paleontology and definition of the name Mammalia. Syst.
879 Biol. 41:372–378.
880 Schander C., Thollesson M. 1995. Phylogenetic taxonomy-some comments. Zool. Scr. 24:263–
881 268.
882 Senderov V., Simov K., Franz N., Stoev P., Catapano T., Agosti D., Sautter G., Morris R.A., Penev
883 L. 2018. OpenBiodiv-O: ontology of the OpenBiodiv knowledge management system. J.
884 Biomed. Semantics 9:5.
885 Schoch R.R. 2013. The evolution of major temnospondyl clades: an inclusive phylogenetic
886 analysis. J. Syst. Palaeontol. 11:673–705.
887 Schuh R.T. 2003. The Linnaean system and its 250-year persistence. Bot. Rev. 69:59–78.
888 Sereno P.C. 1999. Definitions in Phylogenetic Taxonomy: Critique and Rationale. Syst. Biol.
889 48:329–351.
890 Sereno P.C. 2005. The Logical Basis of Phylogenetic Taxonomy. Syst. Biol. 54:595–619.
891 Sereno P.C., McAllister S., Brusatte S.L. 2005. TaxonSearch: a relational database for
892 suprageneric taxa and phylogenetic definitions. Phyloinformatics 8:1–21.
893 Sferco E., López-Arbarello A., Báez A.M. 2015. Phylogenetic relationships of †*Luisiella feruglioi*
894 (Bordas) and the recognition of a new clade of freshwater teleosts from the Jurassic of
895 Gondwana. BMC Evol. Biol. 15:268.
896 Smith B., Ashburner M., Rosse C., Bard J., Bug W., Ceusters W., Goldberg L., Eilbeck K., Ireland
897 A., Mungall C., Leontis N., Rocca-Serra P., Ruttenberg A., Sansone S.-A., Scheuermann R.,

898 Shah N., Whetzel P., Lewis S. 2007. The OBO Foundry: coordinated evolution of ontologies
899 to support biomedical data integration. Nat. Biotechnol. 25:1251–1255.
900 Smith S.A., Beaulieu J.M., Stamatakis A., Donoghue M.J. 2011. Understanding angiosperm
901 diversification using small and large phylogenetic trees. Am. J. Bot. 98:404–414.
902 Smith S.A. Brown J.W. 2018. Constructing a broadly inclusive seed plant phylogeny. Am. J. Bot.
903 105: 302-314.
904 Soltis D.E., Smith S.A., Cellinese N., Wurdack K.J., Tank D.C., Brockington S.F., Refulio-
905 Rodriguez N.F., Walker J.B., Moore M.J., Carlsward B.S., Bell C.D., Latvis M., Crawley S.,
906 Black C., Diouf D., Xi Z., Rushworth C.A., Gitzendanner M.A., Sytsma K.J., Qiu Y.-L., Hilu
907 K.W., Davis C.C., Sanderson M.J., Beaman R.S., Olmstead R.G., Judd W.S., Donoghue M.J.,
908 Soltis P.S. 2011. Angiosperm phylogeny: 17 genes, 640 taxa. Am. J. Bot. 98:704–730.
909 Spatafora J.W., Chang Y., Benny G.L., Lazarus K., Smith M.E., Berbee M.L., Bonito G., Corradi
910 N., Grigoriev,I., Gryganskyi A., James T.Y., O'Donnell K., Roberson R.W., Taylor T.N.,
911 Uehling J., Vilgalys R., White M.M., Stajich J.E. 2016. A phylum-level phylogenetic
912 classification of zygomycete fungi based on genome-scale data. Mycologia, 108:1028–1046.
913 Sterli J., Pol D., Laurin M. 2013. Incorporating phylogenetic uncertainty on phylogeny-based
914 palaeontological dating and the timing of turtle diversification. Cladistics 29:233-246.
915 Stevens P.F. 2006. An end to all things?—plants and their names. Aust. Syst. Bot. 19:115–133.
916 Sundberg P., Pleijel F. 1994. Phylogenetic classification and the definition of taxon names. Zool.
917 Scr. 23:19–25.
918 Thau D., Ludäscher B. 2007. Reasoning about taxonomies in first-order logic. Ecol. Inf. 2:195–
919 209.
920 Thau D., Bowers S., Ludäscher B. 2008. Merging taxonomies under RCC-5 algebraic articulations.
921 Proceedings of the 2nd international workshop on Ontologies and information systems for the
922 semantic web. p. 47–54 doi:10.1145/1458484.1458492
923 Thessen A.E., Bunker D.E., Buttigieg P.L., Cooper L.D., Dahdul W.M., Domisch S., Franz N.M.,
924 Jaiswal P., Lawrence-Dill C.J., Midford P.E., Mungall C.J., Ramírez M.J., Specht C.D., Vogt
925 L., Vos R.A., Walls R.L., White J.W., Zhang G., Deans A.R., Huala E., Lewis S.E., Mabee
926 P.M. 2015. Emerging semantics to link phenotype and environment. PeerJ. 3:e1470.
927 Torres-Carvajal O., Mafla-Endara P. 2013. Evolutionary history of Andean Pholidobolus and
928 Macropholidus (Squamata: Gymnophthalmidae) lizards. Mol. Phylogenet. Evol. 68:212–217.

Vision T., Blake J., Lapp H., Mabee P., Westerfield M. 2011. Similarity between semantic description sets: addressing needs beyond data integration. In: Kauppinen T., Pouchard L.C., Keßler C., editors. Proceedings of the First International Workshop on Linked Science (LISC 2011). Bonn (Germany): CEUR Workshop Proceedings.

Vogt L. 2009. The future role of bio-ontologies for developing a general data standard in biology: chance and challenge for zoo-morphology. Zoomorphology 128:201–217.

W3C OWL Working Group. 2012. OWL 2 Web Ontology Language Document Overview (second edition). https://www.w3.org/TR/owl2-overview.

Washington N.L., Haendel M.A., Mungall C.J., Ashburner M., Westerfield M., Lewis S.E. 2009. Linking human diseases to animal models using ontology-based phenotype annotation. PLoS Biol. 7:e1000247. 10.1371/journal.pbio.1000247

Wojciechowski M.F. 2013. Towards a new classification of Leguminosae: Naming clades using non-Linnaean phylogenetic nomenclature. S. Afr. J. Bot. 89:85–93.

Wright D.F., Ausich W.I., Cole S.R., Peter M.E., Rhenberg E.C. 2017. Phylogenetic taxonomy and classification of the Crinoidea (Echinodermata). J. Paleontol. 91:829–846.

Wyss A.R., Meng J. 1996. Application of phylogenetic taxonomy to poorly resolved crown clades: a stem-modified node-based definition of Rodentia. Syst. Biol. 45:559–568.

Zimmermann W. 1931 (1937). Arbeitsweise der botanischen Phylogenetik und anderer Gruppierungswissenschaften. In: Abderhalden, E., editor. Handbuch der biologischen Arbeitsmethoden (Abt. 3, 2, Teil 9). Berlin (Germany): Urban & Schwarzenberg. p. 941–1053.

Zimmermann W. 1934. Research on phylogeny of species and of single characters. Am. Nat. 68:381–384.

Zimmermann W. 1943. Die Methoden der Phylogenetik. In: Heberer, G., editor. Die Evolution der Organismen 1. Aufl. G. Jena (Germany): Fischer. p. 20–56.

Figure captions

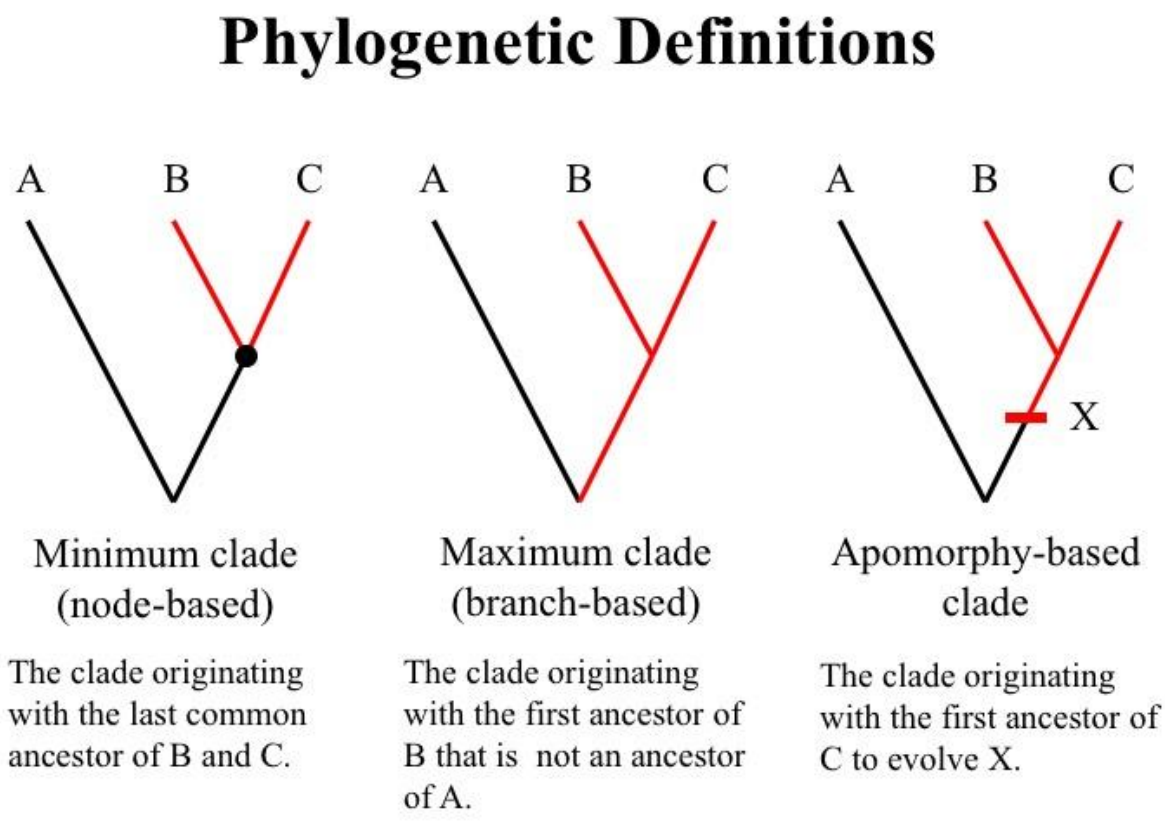
Figure 1. The three basic clade definitions.

Figure 2. Phylogeny of Asterales showing the clade Campanulaceae with all five lineages, the Rousseaceae, and other related lineages.

Figure 3. Upper half: phylogeny of Campanuloideae redrawn from Crowl et al. (2016) showing the polyphyly of *Campanula* (lineages in blue). Lower half: Distribution of *Campanula* as retrieved from a GBIF query.

991 Figure 1.

992



993

994

995

996

997

998

999

1000

1001

1002

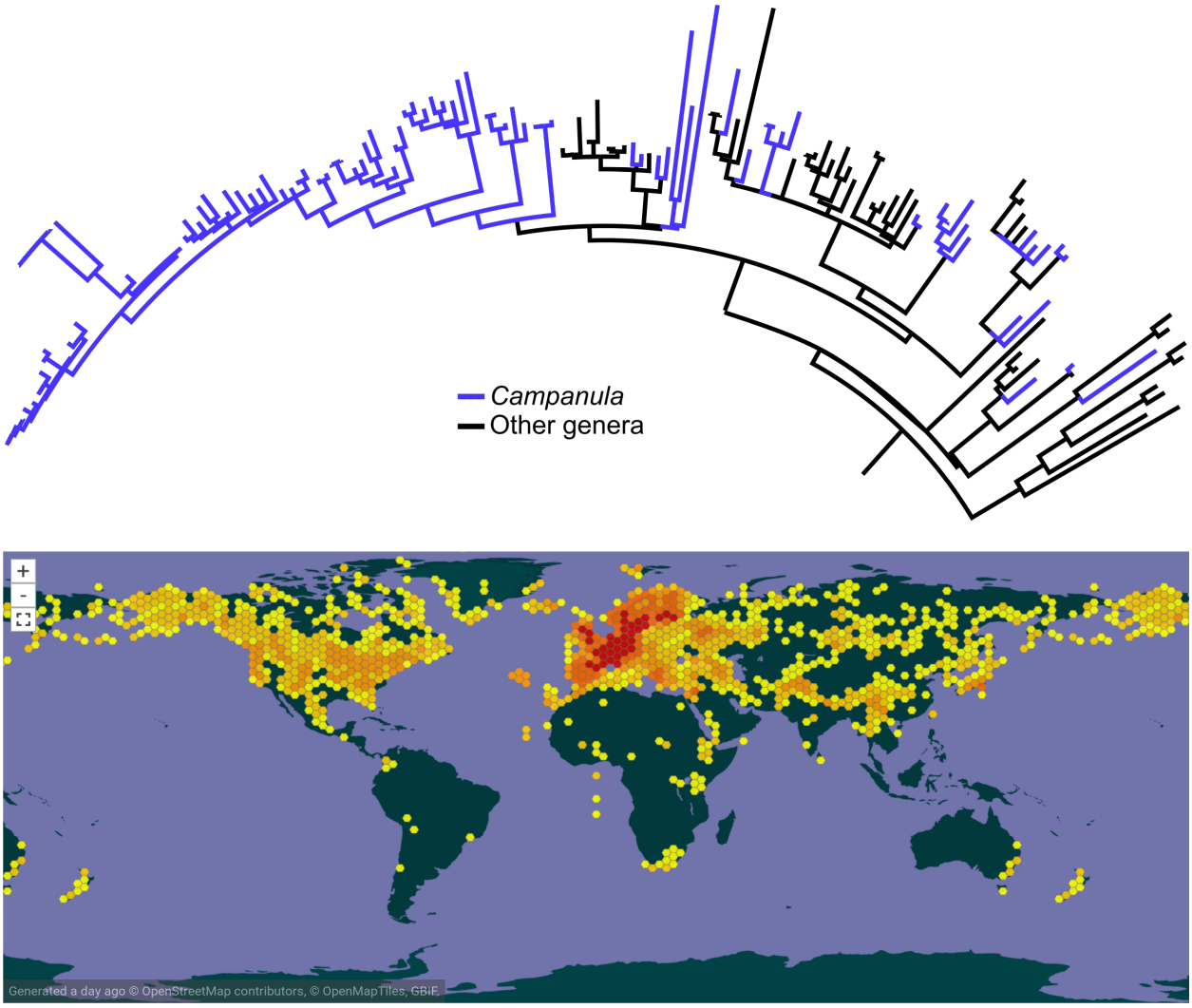
1003

1004

1005

1006 Figure 2.

1007



1008

1009

1010

1011

1012

1013

1014

1015

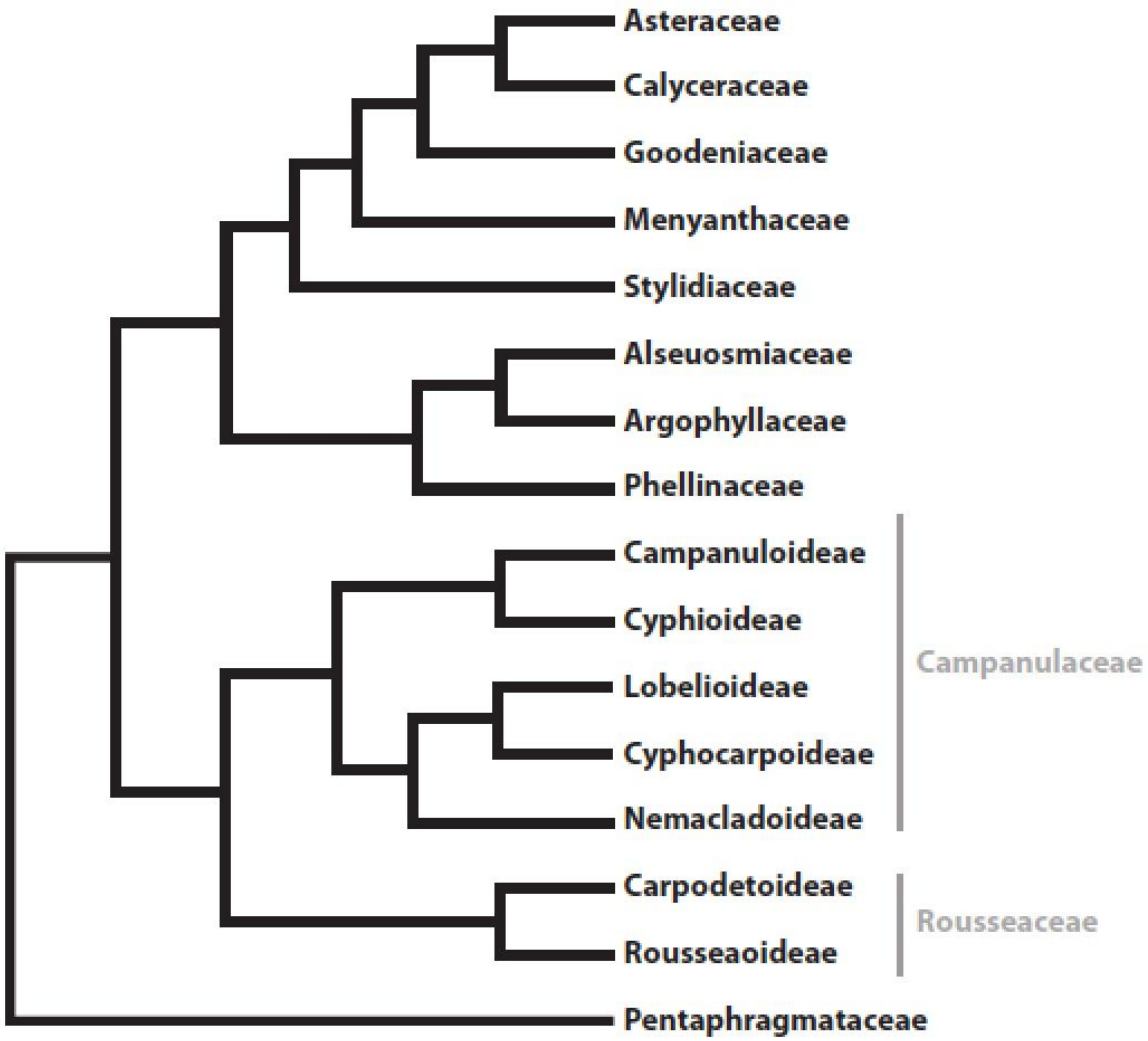
1016

1017

1018

1019 Figure 3.

1020



1021