4 CONSTRUCTING TAXON CONCEPTS

> An organizational scheme might choose any features at all of the writings as the basis for assignment to position: size of copies, weight of copies, security classification, legal or social function, religious orientation. But of the familiar sorts of organizational scheme, those resembling The Catalog in assigning positions on the basis of subject matter are perhaps the most interesting and problematic.
>
> —PATRICK WILSON
>
> *Two Kinds of Power: An Essay on Bibliographical Control* (1968, 69)

ON CONCEPTS AND ARTIFICIALITY

One of the greatest powers of the classifier is, quite literally, to make concepts and entities *exist* within a system—to instantiate concepts as tools for description and conduits for information access. To classify is to have the *power to* create a concept within a system, as well as the *power over* how people both conceptualize that system and subsequently use that system to access resources. Whether the dingo exists in a classification as a genuine species of its own dictates the way Australian politicians and governments can act in response to whether it is seen as a pest in the country or not. But what, exactly, does it mean when I say, "Classifications bring entities into existence?" Surely, the organism we associate with being dingo-like existed before the name *Canis lupus dingo* or *Canis familiaris dingo* was formally entered as a classification. Just as entities in the natural world existed before Linnaeus articulated the formal rules for the application of binomial nomenclature.

The point of this question reaches to the heart of what is meant when I say that classifications are constructed: all the concepts and entities that we include in classifications are socially produced, even if their articulation is part of an empirically based scientific process. To say that there are no natural categories is to claim that all attempts to fragment the world into disparate entities is, and always will be, an artificial activity. Classification may be a natural human inclination, yes, but the classes we produce are a product of our own cognitive and disciplinary limitations, spatiotemporally contingent interpretations, and methodological possibilities. "Living systems must categorize. Since we are neural beings, our categories are formed through our embodiment. What that means is that categories are a part of our experience! . . . We cannot . . . 'get beyond' our categories and have a purely uncategorized and unconceptualized experience" (Lakoff and Johnson 2010, 19). Categorization and classification are always performed relative to one's self and the society in which they are embedded. Even the "natural" cannot exist on its own, for to understand what is natural we must relate it to what is not natural—namely, everything that humanity has created in an artificial sense, such as cities, towns, and other human-made environs. Additionally, these classification systems inherit the just and unjust elements of that same society from which they emerge. There is no foundational rubric on how to distinguish one species of giraffe from another, just as there are no preestablished guidelines on how we should categorize one's sexual preferences, ethnicity, gender identity, or, far more innocuously, various shades of blue or red. How we set the boundaries for these classes is a sociocultural issue.

Importantly, a crucial historical distinction has been made between what biodiversity taxonomists have considered "natural" and "artificial" systems. So, while much of this discussion is in service to IS scholars perhaps unacquainted with biological taxonomic history, the reality is that working biological taxonomists are well aware of the artificiality of their constructions. According to Staffan Müller-Wille, the distinction between natural and artificial systems first arose when the distinction was identified by Carl Linnaeus (Müller-Wille 2013), though Phillip Sloan (1972) traces the

genesis of this juxtaposition to the seventeenth century, especially by way of the writings of John Ray, Joseph Pitton de Tounefort, and August Bachmann. In a general sense, the articulation of a natural system has, as noted by Phillip Sloan, attempted to devise a method of taxonomic construction that "neither reflects whimsy of the taxonomist, nor represents simply a utilitarian cataloguing device like the Dewey Decimal system" (1972, 2). How any one scientist defines what constitutes an appropriate method to devise a natural system has varied by time period—conservatively ranging from methods that identify a single character, or finely limited set of characters, as viable candidates to inform the production of a natural system (the fruit-producing parts of plants, for example) to whether a great many characters should be considered for the construction of such a system (such as in phenetical taxonomy). As Müller-Wille notes, natural systems, at least since the time of Linnaeus, can more broadly be defined as those that "conform to intuitions about group membership that are based on overall resemblance" (Müller-Wille 2013, 310), using a variety of characteristics to devise a comprehensive comparative study. The goal for defining a natural system is, at least in part, to locate an inherent natural, intrinsic order by empirical means (Lefèvre, 193), however that 'intrinsic nature' might be defined at a given time period. As also pointed out by Sloan, how this natural ordering has been interpreted (that is to say, what it has represented) has differed based on historical period. For pre-Darwinian taxonomists, a natural system was one that mirrored the "true essence" of the natural world (Sloan 1972, 2), inclusive of arrangements such as the Great Chain of Being, which held God, angels, and humans at the apex of the organism- and mineral-based hierarchy. In a post-Darwinistic world, a natural system came to reflect a phylogenetic framework that was most likely accurate to evolutionary history—a belief that still is, in many ways, latent in some of the discourse surrounding phylogenetic classificatory arrangements (Doolittle 1999a; Liu, He, and Schneider 2014).

As noted by Stefan Müller-Wille (2013), it was Linnaeus who first used the terms natural and artificial to distinguish these concepts; to foist this distinction squarely over disagreements prior to Linnaeus's time is to

fall into the danger of anachronism and oversimplification of the disagreement. Linnaeus believed that his own system, as well as all other taxonomies devised up to that point, were definitionally artificial, and that to have any chance at devising a natural system would entail a near full catalogue of organism characteristics (Müller-Wille 2013)—a feat unattainable in the eighteenth century and perhaps debatably an impossible task even in the present or at any point in the future. Artificial systems, then, are context-dependent taxonomies, based on the local and practical conditions under which these taxonomies are produced. These classifications are based on limited character comparisons and necessarily produce pragmatic classifications. In an ideal world, a natural system would require no great revisions upon the addition of a new organism. After all, a natural system would be comprehensive enough to fit all current and future organismic discoveries. Artificial systems, however, are often dependent on the select characteristics and subject to great change upon the addition of new species—for example, particularly when new specimens introduce ambiguity with respect to the characteristics used as the principal divisions for a taxonomic hierarchy (Müller-Wille 2013, 311).

So, it is one thing to speak of the arbitrary nature of classification in a general sense—as in Sloan's indication that the Dewey Decimal Classification is utilitarian—but it is quite another to transport those assumptions full-scale to the biodiversity classification world. My use of the word artificial may gesture to this historical contention, but in general, my aim is to take this artificiality as a given (as most all biodiversity taxonomists do) and to better understand how the conditions under which classifications are constructed produce certain modes of systemic imbalances of power. We know that taxonomic practice is anything but arbitrary and that the construction of biological classifications is based on careful scientific analysis and methodological inference. Much like the concept of literary warrant in information studies (Beghtol 1986), the construction of a new class (bibliographical subject or taxon, for example) requires the establishment of evidence that justifies the need for a new class. To say they are artificial is not a slight, but rather a reality of classificatory construction.

SPECIES CONCEPTS

With this distinction established, two terms merit definition before we proceed onward: species concepts and species taxon concepts, which Walter Bock (2004) notes are too often confused in biological and systematic literature. First, to the species concept: as defined by Bock (2004), the most widely recognized is the biological species concept (BSC) or genetical species concept (GSC). The GSC is defined as a "group of actually or potentially interbreeding populations which are *genetically* isolated in nature from such other groups" (Bock 2004, 180; emphasis original) The BSC, originally identified by Ernest Mayr, is slightly narrower, noting that a species comprises a population that is reproductively isolated. But as Bock has noted, some species can interbreed without producing viable offspring; for example, a horse and a donkey can interbreed to produce a mule, but the mule is infertile because of chromosomal differences between the two parent species. It is the case that genetic isolation is concomitant with reproductive isolation, but in this case, the primary determining factor is genetic. It is not always the case that reproductive isolation equals genetic isolation. In addition to these concepts, as summarized by Marc Ereshefsky (Ereshefsky 2007, chap. 2), quite a few other species concepts are also in circulation—some say upward of twenty-two to twenty-six, in fact (Wilkins 2011). Each species concept has its own theoretical approach to defining how organisms can be meaningfully and functionally grouped into evolutionary units. For example, a phenetic species concept is one that defines a group of organisms based on their resemblances with each other. This definition is operational "based only on observable facts of similarity and discontinuity" (Winston 1999, 44). Here, one might say, after a close morphological examination of our aforementioned giraffes, they are one species because they are similar enough to one another that such a determination makes good empirical sense. Whether they can reproduce with each other might, for some, solidify such a categorization, but reproduction would not make or break a phenetic determination. Another species concept is the ecological species concept, which defines a species based on the environmental niche they fill and the extent to which a group utilizes a

particular set of natural resources in an ecologically efficient way. Such an approach prioritizes environmental forces and conditions as the stabilizing force for a species, over and above reproductive isolation (Ereshefsky 2007, 87–90). In this way, the operative variable in the ecological species concept is the environment, even if interbreeding is also a quality of a group.

It is imperative to note that the species concept is a term that applies to the domain of evolutionary biology; its articulation is separate from the practices that taxonomists undergo creating species categories (a given level in the Linnaean hierarchy) and the species taxon, which is delimited (described) (Bock 2004). A working taxonomist may adopt any one species concept, and the subsequent articulation of species categories and species taxa will be influenced by that choice. It is often the case that one's research context determines the best species concept to implement in practice. As Kevin De Queiroz indicates, "Various properties [of species concepts] are of greatest interest to different subgroups of biologists. For example, reproductive incompatibilities are of central importance to biologists who study hybrid zones, niche differences are paramount for ecologists, and diagnosability and monophyly are fundamental for systematists" (2007, 880). Regardless of which species concept is adopted, the concept must then be consistently applied to all species taxa within a given hierarchy (Bock 2004, 185). Adopting different species concepts in the same taxonomic construction will inevitably lead to contradictory, and possibly incompatible, species categories and species taxon (De Queiroz 2007, 879–880). For one, some concepts facilitate the creation of more taxa (i.e., genetical or phylogenetical species concept; colloquially, a "splitter"), whereas another species concept might create taxa inclusive of more organisms, and thus fewer taxa (i.e., biological species concept; a "lumper") (2007, 879–880).

TAXON CONCEPTS

Taxon categories are the basic building blocks for the Linnaean hierarchy and species taxa in particular are the fundamental and foundational taxon concept on which all other taxa and hierarchy categories are defined (De Queiroz 2007, 181). Species concepts (as described above) inform the

conceptualization of taxon categories in general, and species taxa in specific, by influencing how a particular species taxon is delimited (bounded), described, and related to other taxa. Taxon concepts are empirical formulations in the strongest sense of the word and function as any other scientific opinion might—as a hypothesis, subject to future testing and revision (Wiley and Lieberman 2011, 29). For example, the articulation of the species taxon concept and the delimitation for *Ursus arctos* (the concept commonly known as the brown or grizzly bear) can be challenged at any point by reformulating the taxonomic circumscription and concept through a new published delineation. Summed up by Sterner, Witteveen, and Franz,

> Unlike the choice of a species concept, which given the current state of the species debate can perhaps be treated as somewhat arbitrary, taxonomic concepts have an empirical status as scientific hypotheses. To see this, consider the difference between identifying what a given name refers to with and without specifying a taxonomic concept. It is easy and uncontentious to say, "This taxonomic name refers to the biological entity that includes this type specimen." While correct as a matter of principle, a statement of this kind fails to communicate anything about the present state of knowledge about the relevant species. It is much harder epistemically to accurately identify and agree on which organisms other than the type specimen are also members of the designated species. (2020, 8)

Taxon concepts are *definitional* concepts, in that a taxon represents a particular taxonomist's informed opinion about what constitutes a bounded population of real-life organismic units as observed in some natural environment (De Queiroz 2007, 182), ultimately represented by species nomenclature.

A taxon concept is complex, represented as it is within a computational system by a name that references numerous external sources, including the type specimen that anchors the name (Patterson et al. 2006). In Ronald Day's terms, "the name points to the object and the name reflects the networks in which the object first appears as a named thing" (Briet 1951, 49). The ties that bind these disparate sources of evidence, however, are shifting and difficult to trace as changes progress over time, such as a revision to, or a shift in the articulation of, a particular taxon concept. A species taxon concept in an operational sense is an accumulation of a

name string in a technical system, as well as the compilation of referenced evidentiary objects that make that name valid as part of scientific discourse (the specimens used for description, including a type), as well as a publication that delimits and describes the taxon. Without this "trinity" in place (the type, the description or delimitation, and the taxon's stable nomenclature) a concept fails to make it into formal scientific discourse. Typification involves the identification of a type specimen—an individual or group of specimens. Types are typically held in museum repositories and act as the material ground for and the name-bearer of the species (Daston 2004). Description occurs through a publication act and includes a species diagnosis, a description, a taxonomic discussion, and, potentially, notes about a species taxon's ecological characteristics, geographic distribution, and other matters (Winston 1999). Instantiating a species taxon also involves attaching said delimitation to a name, which is of special interest in this discussion since names are the primary collocating mechanism for data about a species once taxon concepts are brought into a taxonomic space.

The Type Specimen

Most have a general idea that one of the primary documents of biological taxonomy is the type specimen—the one and most obvious physical thing that is "attached" to a name, that creates the species "out there." Michael Buckland's ruminations on museum specimens-as-documents (2017) and Suzanne Briet's (1951) trope of the antelope have permeated IS and documentation studies such that the type is often seen as the most vital taxon-representing object. Briet is well known for popularizing one of the most well-known definitions of the "document" in the field of IS and documentation studies: "a document is a proof in support of a fact" (1951, 9). To illustrate the complicated indexical relationship between documents and their referent, she uses the example of an antelope in the wild, which is then captured and brought back to Europe, with its ultimate fate being "stuffed and preserved in a museum" (1951). In the wild, the antelope is merely a natural object; but in a botanical garden in Europe or as a specimen in a museum, however, the once free and live antelope is now a document that indexes the "kind" of the antelope that lives free. Briet also describes many

other documents that are created in this process (articles about the antelope, photographs, etc.). Many scholars in IS and documentation studies use these distinctions as a jumping-off point for theoretical examinations. The truth is that the type specimen is not related to the taxon concept as directly as we might think—and certainly not as neat and straightforward as Briet's exemplar. But this assumption subsists for good reason. After all, when establishing whether a given organism is a new species or a new discovery, one often, as Judith E. Winston states, must look to this physical object as evidence for taxa determinations (1999, 96). But while types do occasionally settle taxonomic disputes (Godfray 2007) and can also act as the voucher specimens on which some molecular examination of species are based (Seberg et al. 2016; Smithsonian Institution 2017), they are not always determinative. The most important function for type specimens remains to anchor and regulate nomenclature by preventing the chaotic inflation of names (Witteveen 2015, 570; Winston 1999, chap. 9). With this stability, the name, and its associated taxon concept, can then be argued in the space of scientific activity.

When Linnaeus began using binomial nomenclature, he based his description on specimen samples, but the formal identification of types was not yet a global standard (Ereshefsky 2007). Not until the late nineteenth century did typification enter common usage, followed thereafter by its formal integration into the codes of nomenclature (Daston 2004, 154–158). Before such rules were established in codes, the assumption that the type *be* typical led to an inflation of the application of types, as well the frequent replacement of types for newer exemplars that more adequately matched phenotypic observations (Hull 1988, 498). New names arose everywhere, many for the same species, leading taxonomists to mandate the use of types. Only one type could be applied to one name; by doing so, the production of names became more centralized and controlled (1988, 498), focusing efforts more on the process of taxon circumscription. The result of this codification is that, regardless of how a taxon is described over time, it remains what Joeri Witteveen calls a "necessary truth that the taxon's type specimen falls within its boundaries" (2015, 569).

And while the type is presumed to be "in one's mind" while describing and delimiting a species taxon, in reality, the type is not a "typical"

representation of the taxon it is eventually associated with. One particular individual of a species can potentially vary descriptively from any other individual—males might look different from female specimens, and specimens might look different based on life-stage differences or natural variability. In general, best practices have been established for identifying a type. Ideally, when possible, a type should be selected from a robust and large population where variation is likely most ideal (Winston 1999, 177). In most cases, one specimen is identified as the primary name-bearing document; this specimen is called the holotype, which means a name is *always* directly tied to that one specimen. In addition to the holotype, paratypes can be identified, which are specimen references for description and delineation, but which are not in any way connected to the instantiation of a name (ICZN 1999, sec. 73). If many types of equal stature are used to describe a species, they are collectively known as syntypes, though this practice is certainly less favored within institutions, given the complexity of preserving types and managing their collocation in a collection.

Types can be of full or partial organisms, or some fossilized remains or impression of a previously extant species (ICZN 1999, sec. 72.5). As Rogers et al. note, "specimens are testable, tangible, and verifiable data sources" (2017, 456). Photographs proper cannot generally serve as the type without a great deal of scrutiny, as least not in the ICZN (1999, sec. 72.5.6). Though it is possible to base a species description on a photograph, description, or illustration of a type specimen, the "the name-bearing type is considered to be the specimen(s) illustrated or described" (ICZN 2018), not the image itself.[1] Rules in the code allow for the application of a new type when one is lost—called a neotype. However, as the hunting and capturing of certain species is outlawed, certain exceptions to this rule have been articulated. Fervent discussion has ensued regarding whether or not high-resolution photographs are a proper stand-in for physical specimens (Garraffoni and Freitas 2017; Rogers et al. 2017). One fear is that such an approach will potentially increase typification without preserved material (Krell and Marshall 2017), proving counterproductive to the purposes of types in nomenclatures rules. Another is that this creates an "inherent data deficiency problem: images cannot contain all morphological data or

any genetic data possessed by actual specimens and therefore have limited utility in the face of growing knowledge" (Rogers et al. 2017, 455). As it stands, the practice is highly discouraged, at least until the scientific community agrees on a set standard, though it does seem that the issue will necessarily be reckoned with as the rate of extinction increases for species across the planet.

These type specimens stand at the center of sound and verified taxonomic work, playing a major role in correctly (as in, according to stated rules) assigning names to taxa. This is one reason museums take their conservation very seriously (figure 4.1). In a rather traditional sense, types are reference objects used by scientific professionals to initially describe taxon concepts in publications, which are required to formally instantiate a nomenclature act. As Timothy Utteridge, head of identification and naming and senior research leader at Royal Botanic Gardens, Kew, explained, subsequent taxonomists then use type specimens to confirm the connections among the publication, the type, and the actual population in the field when appropriate and necessary, especially when specimens are limited in the wild (interview 2016). The original taxon concept circumscription is,

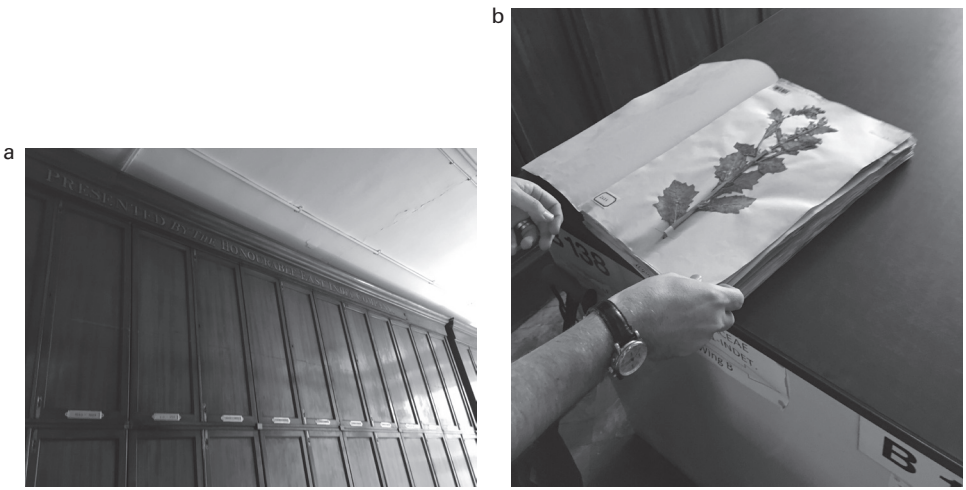


Figure 4.1
(Left) Original East India Company type specimen cabinets. (Right) A type specimen folder from the East India Company cabinet. Royal Botanic Gardens, Kew. Photo by author.

by default, connected to the type specimen insofar as it remains tethered to the name that represents the concept—any emendations to a concept must negotiate this material ground and the name it represents.

Publication and Naming

Type specimens bear the name and control the inflation of nomenclature, but the publication itself instantiates the name as a valid token to be used in scientific discourse. Publications describing a new taxon concept provide many qualitative data points of interest, including the circumscription of the species taxon, known synonymic nomenclatural variants, and type material identified, as well as a host of other potential descriptive information points (Winston 1999, pt. 3). Information contained in the publication is of central importance in the conceptualization of a new taxon, providing the necessary information for future scientists to confirm or refute a particular concept. Careful attention is paid to the characteristics that differentiate a new taxon from those that are closely related, a process known as diagnosis. Diagnosis identifies how the shared traits for taxa might differ and how taxa "differ completely" (Winston 1999, 190). In the example cited by Winston (1999, 117), the taxon concept for *Batillipes gilmartini* is differentiated from associated taxa by a "distinctly different caudal spine and the phylogenetically significant dorsal plates" (McGinty 1969).

In order for names to enter the communication stream of scientific discourse, they need to be within the domain of public knowledge (Wilson 1977), such that any and all elements related to a taxon concept can be accessed for reference, confirmation, refutation, and revision. Not just any publication venue will do. Within the *International Code for Zoological Nomenclature*, for example, for a species name to qualify as valid, publications must be permanent and freely accessible (ICZN 1999, sec. 8.1). The introduction and proliferation of digital media has required the caveat that a publication is inclusive of electronic formats so long as the content of the publication is fixed in some manner (1999, sec. 8.1.3). Good taxonomic work rests on Patrick Wilson's notion of the complete library, a documentary repository that contains all relevant published material for biodiversity work (Wilson 1977, 87).

The date of a particular published circumscription is crucial, as this imprinted date on the document is used to assess the priority of any given name over another—a critical aspect of nomenclatural standardization (Bowker 2008, 159). The identification of a publication date, however, is not always as straightforward as it may seem—a fact perhaps immediately understood by professional cataloguers and bibliographers. In conversation with zoologists at the Smithsonian National Museum of Natural History (NMNH), we learn that finding an accurate date for older publications is often difficult. Many of the zoological journals held in the NMNH dating from the early twentieth century, for example, have an accession date stamp (the date a journal was added to the NMNH collection), but do not have an imprinted published date on the issue—either by design or by a missing copyright page. Sleuthing, then, becomes a paramount task for nomenclature specialists assessing name priority. Interpolating dates from within the historical record is not uncommon, especially before the implementation of codes that required publication. Timothy Utteridge describes one such example:

> The genus I work on, Maesa that was recognized by, I think, a Danish guy . . . [on] an expedition into Saudi Arabia. And then at the same time, the Forsters went around, I think with Cook, and came back with another [sample] and they called it Baeobotrys. [The Fosters then] published this [description] and it's exactly the same time as Maesa. But the only way [we know] which one takes priority is that someone worked out when their ship landed in Portsmouth, how long it would take them to dock, how long the post carriage took from Portsmouth to London, [and] how long the editor would have taken to write it up. So, they've [assessed], to the day, which one has priority and it came out [that it was Maesa]. (Interview 2016)

These publications provide the documentary (or bibliographical) warrant necessary to prioritize one name string over a constellation of other possible name strings. The function of a warrant-based system, as Claire Beghtol (1986, 110) makes clear, is to provide taxonomists a way to understand the evolution of name types—and their surrounding publications—over a broad period of time.

INSTANTIATIVE POWER

> Naming information is the term I use for creating document surrogates. . . . I choose the word "naming" because it connotes the power of controlling subject representation and, therefore, access. . . . Theories, models, and descriptions are elaborated names. In these acts of naming, the scientist simultaneously constructs and contains nature.
>
> —HOPE A. OLSON
> *The Power to Name: Locating the Limits of Subject Representation in Libraries* (2002)

Biodiversity and IS work both are heavily involved in the production of names and their subsequent control—names that ultimately impact the lived natural and social worlds. I think it is relatively easy to get lost in the procedural aspects of nomenclature production and taxon concept construction, particularly when the process is so codified by way of rules of nomenclature, delineation, and description. The impact of this process, though, should not be underestimated, as it lays out the processes by which formal taxon units are created, on which all taxonomic work rests. Biological names instantiate concepts, which can then be applied and operationalized in the process of taxonomy and class construction for real world entities. To have this kind of instantiative power is to maintain control over what can and cannot be entered in the intellectual field of record, where species are formally articulated as existent or not within a host of biodiversity databases, records, and documents. In the case of our imperiled dingo, the stakes could not be higher when we consider whether a certain formal infrastructure (whether the Catalogue of Life or some other database) adopts the formal designation of *Canis lupus dingo* or *Canis familiaris dingo* as the prioritized name and the interpretation of its associated taxon concept. And certainly, the best taxon concept construction work requires that the full historical trajectory of any given concept be available to make the best determination based on the preponderance of scientific evidence.

Within the bibliographical realm, name production and instantiation have a very particular application as well, that serves as an analogue for change of a different sort. In Patrick Wilson's *Two Kinds of Power*,

instantiation is understood as the production of "instances of particular patterns, or types" (1968, 7). In Wilson's context, an instance of a work, for example, is the physical production, performance, or exemplification of some abstract creative endeavor—a book, a play, a song, or the like. Work-instantiation theory, then, looks to formulate formal relations between a work and its copies (see for example, Smiraglia 2001, 2005). Textual criticism and textual bibliography (Gaskell 2007, 313–360), for example, are domains of study that examine the genealogy and editing of texts over time. Or, put another way, these areas follow the production and alteration of work-instances over time. In textual bibliography, for example, work-instance variations might be "compared" with what is called an "ideal copy." An ideal copy, as espoused by Fredson Bowers, is an "ideally perfect" copy of a text as the printer or publisher (and, I might add, the author) originally intended it (1994, 113). From this idealized version of the work, variations and errors introduced during publishing and production can be identified and documented against. G. Thomas Tanselle's (1980) version of ideal copy is a bit more materially focused and composite-based than Bowers's, given that the "reconstruction" of an ideal copy "encompasses all states of an impression or issue, whether they result from design or accident" (1980, 46). As Tanselle continues, "the '*ideal copy*' is central to descriptive bibliography, because it is the element that distinguishes bibliographic description from cataloguing: whereas a catalogue entry, regardless of its level of detail, exists to record a particular copy, a bibliographic description [centrally using the concept of the ideal copy] aims to provide a standard against which individual copies can be measured" (1980, 21; emphasis original). The ideality presupposes a printer's intended perfect instance of a text, which means that, in practice, these ideal copies are more often than not abstract, as this "ideality" is not attainable.

Much like taxon concepts, works and texts are also constantly shifting and evolving along temporal and spatial lines—albeit on very different terms. The identity conditions of a work change over time depending on how its work-instances persist (materially), are reformulated (republished or edited), emended (critical editions, adaptations, and the like), or reinterpreted. Much like Wilson's notion of the subject, the line between one work

and another is infinitesimally graded, requiring the articulation of artificial boundaries to differentiate them. Within information systems, such as a library catalogue for example, the challenge is to draw lines between various instances of works so that the significant differences between each entity can aid in the selection of appropriate resources. In some ways, however, the decision-making process for bibliographers and cataloguers is easy compared to that of the taxonomist studying the minute differentiation of concept change.

On Fish and Control

Even with the concept of nomenclatural priority in effect, tracing the historical records associated with a name is an onerous task. Imagine if name inflation wasn't controlled, how much more difficulty would be encountered as the name disambiguation process played out? Inevitably, if one is to understand the full context of any given name and associated taxon concept, one must take at face value that the circumscription and delineation of the taxon is as fully represented and articulated in a given publication as possible, and that the appropriate references and taxonomic determinations have been documented. And a crucial part of making taxon concept determinations is being able to represent and usefully follow the nuanced changes for that proposed taxon group over a broad period of time. Complicating matters, taxon concepts are anything but stable, particularly in some taxonomic groups. As noted by Nico Franz (2010), only approximately 55 percent of the valid concepts in "eight succeeding classifications of North American vascular plants from 1933–2006" remained stable (2010, 49). The diagnosis section for a new species taxon circumscription, for example, requires taxonomists to examine publications and samples for all closely related species, such that distinctions can merit its inclusion as a new concept. Similarly, revising a taxon requires that one pay special attention to how a particular taxon has permutated over time. The publication of a new or revised species taxon most often also requires a taxonomic discussion section, which outlines the logical reasoning of the taxonomic assessment (Winston 1999, chap. 12). Careful examination of literature (often mostly accessible through nomenclature tokens in a database) is

necessary to properly and ethically articulate new grounds for a taxonomic revision. Quite simply, if taxon concepts-as-names are not readily available, and mapped, the practice of taxonomy comes to a virtual standstill—or worse yet, prior mistakes or observations risk being duplicated or ignored. Preserving the historical continuity of taxonomic knowledge is paramount. As we will see, this kind of change is not usually defined by a simple, linear process. To illustrate how complex change becomes in the practical realm, let us look to an example by Richard Pyle (2008), senior curator of ichthyology at the Hawaii Biological Survey of the Bishop Museum.

Imagine two hypothetical species of fish are extracted from a pool of water in the wild believed to be part of the same genus, named as:

Fish 1: *Holocanthus fisheri* (Snyder 1904) [sec.] Snyder 1904[2]
Fish 2: *Holocanthus acanthops* (Norman 1922) [sec.] Norman 1922

Then Jordan comes along and decides that *Holocanthus fisheri* is actually part of another genus, *Xiphypops*, so he renames it with a new combination moving the genus:

Fish 1: *Xiphypops fisheri* (Snyder 1904) [sec.] Jordan 1922
= *Holocanthus fisheri* (Snyder 1904) [sec.] Snyder 1904

But notice that the concept hasn't changed; it has the exact same circumscription (description) as Snyder 1904.

Then imagine a third scientist described another fish from that same pool, and describes the following as part of a new genus:

Fish 3: *Centropyge flavicauda* (Fraser-Brunner 1933) [sec.] Fraser-Brunner 1933

But Fraser-Brunner also thinks that *all of the fish* from this pool are from this same *new* genus, so she decides to move *all* of the others into the same genus as well:

Fish 1: *Centropyge fisheri* (Snyder 1904) [sec.] Fraser-Brunner 1933
= *Xiphypops fisheri* (Snyder 1904) [sec.] Jordan 1922
= *Holocanthus fisheri* (Snyder 1904) [sec.] Snyder 1904

Fish 2: *Centropyge acanthops* (Norman 1922) [sec.] Fraser-Brunner 1933
= *Holocanthus acanthops* (Norman 1922) [sec.] Norman 1922

In his presentation, Pyle continues to describe how yet another individual comes along and decides that Fish 3 is actually a synonym of Fish 1, and proceeds to bring those two species groups under one genus:

Fish 1 and Fish 3: *Centropyge fisheri* (Snyder 1904) [sec.] Pyle 2003
> Fish 1: *Holocanthus fisheri* (Snyder 1904) [sec.] Snyder 1904
> Fish 1: *Xiphypops fisheri* (Snyder 1904) [sec.] Jordan 1922
> Fish 1: *Centropyge fisheri* (Snyder 1904) [sec.] Fraser-Brunner 1933
> Fish 3: *Centropyge flavicauda* (Fraser-Brunner 1933) sec Fraser-Brunner 1933
= Fish 1: *Centropyge fisheri* (Snyder 1904) [sec.] Fraser-Brunner 1933
+Fish 3: *Centropyge flavicauda* (Fraser-Brunner 1933) [sec.] Fraser-Brunner 1933[3]

It was a circumstance such as this that led Pyle to remark jokingly during the 2008 annual Biodiversity Information Standards/Taxonomic Databases Working Group (TDWG) meeting, "Taxonomy is the perpetual classification of mis-named species" (2008). Multiple names can represent the same taxon concept (synonyms); one name can be used for many entirely different concepts (homonyms); and one name can refer to two or more concepts, whose circumscriptions overlap (usually resulting when taxa are split or merged over time) (Remsen 2016).

Herein lies the problem, according to Pyle: "Sometimes the same concept goes by different *legitimate* names and sometimes the same name can refer to different *legitimate* concepts" (2008; emphasis added). In light of this, the historical network of taxon name tokens becomes increasingly difficult to parse and differentiate. As Pyle's example shows, name formulations have the capacity to include semantically embedded metadata (author, date, and so on). But such information is, at best, only occasionally included as part of a name string. Most name-forms do not include these provenance markers, despite calls from taxonomists that they be consistently included (Franz, Peet, and Weakley 2008). Nico Franz has also been

vocal about the importance of computational ontologies to bring order to nomenclatural issues such as those described above (2010). And although these ontologies cannot solve all the problems inherent in nomenclatural control, they certainly have more ability to map the qualitative differences among name tokens in ways more informative than hierarchical presentations or traditional nomenclatures. Relationship indicators such as "synonym of" can produce more meaningful relationships that also have the ability to be mapped over time (Groß, Pruski, and Rahm 2016).

Concept determination and concept change play out in different ways in IS, of course, but concept mapping is equally important in both fields. One analogue is how we go about determining the historical evolution of subject terms in classification systems, and how these studies can illustrate certain cultural and academic trends over time. As Joseph Tennis (2002) relates in reference to his subject ontogeny studies, "The power of a classification scheme to collocate is compromised if we do not account for scheme change" (2012, 1350). Scheme change, like taxonomic change, is useful not only as a document of the classification itself, but also to facilitate information retrieval as subject and access terms evolve—a matter not wholly different from the problems faced with nomenclature in the biodiversity taxonomic world. As noted by Ellen Greenblatt, subject term change can represent the mapping of similar terms in syndetic relationships, such as mapping the terms "lesbian, dykes, and gay womyn" (2010, 212). And, akin to Tennis's study, Greenblatt also notes that obsolete terms must also be mapped onto their contemporary permutations, such as how the "terms *lesbigay* and the more inclusive *lesbigaytr*," once in vogue in the 1990s, have now fallen out of colloquial use (2010, 212). Certainly, as more robust infrastructure increasingly arises to more efficiently represent taxonomic changes over time, IS should be poised to take lessons from taxonomic professionals that deal with such change on an exceedingly more complex level.

CONTROL, OUTLINED

The first step toward nomenclatural control is aggregating all undifferentiated name-forms produced around the world, ranging from those that are

in correct scientific form to vernacular and common names to any possible iteration between these two poles. In this initial stage, names have not yet been disambiguated or validated by nomenclature professionals. Valid names are both well-formed (syntactically) and attend to the nomenclatural rules specified within their particular domain in terms of both syntax and semantics (independent rules exist for the botanical, zoological, and viral taxa, for example).

The Global Names Architecture (GNA) has arisen to serve this vital name-collocating function in the biodiversity world (GNA 2021c). The GNA locates and indexes name instances scraped from across the web and links these name instances directly to their original source (2021b). The GNA is composed of two distinct parts: the Global Names Index (GNI) and the Global Names Usage Bank (GNUB). The first and less curated of the two, the GNI, is a list in the broadest sense, and includes any number of name-forms, including code-compliant scientific names, common names, genetic barcodes, and any other generic species identifiers (GNA 2020a). Any token that points to a species or species concept is included. Because of its catch-all nature, the space is often jokingly referred to as the "dirty bucket" (GNA 2020a). As of April 2021, the GNI data bank contained more than seventeen million name-forms (GNA 2021a). Content for the GNI is variable, unstandardized, and contributed by many organizations and sourced from both online and digitized analog sources, including articles, databases, websites, and online specimen repositories where labels and other museum data are scanned with OCR software for easy ingest. Institutions such the Natural History Museum, London; Royal Botanical Gardens, Kew; and the Smithsonian National Museum of Natural History regularly contribute digitized literature. Given this distributed harvesting methodology, the GNI necessarily contends with hundreds, perhaps thousands, of list sources containing thousands or hundreds of thousands of names, many of which are far less curated than those at large, prosperous institutions such as Kew. Error reconciliation and name disambiguation become the greatest challenge in this environment—a matter we will deal with in just one moment.

Associations

It benefits our narrative to mention that names within the GNI do not just float without reference; they are tied to their documentary context. Associations between a name and a document source become especially important once nomenclature specialists begin validating taxon concepts and assessing the priority of token forms in the pool of potential concepts. But the relationship between a text string in the GNI and its documentary source is relatively flat and provides very little metadata related to the name's full relationship within its source. A mere one-to-one linked relationship makes it difficult for algorithms to link related and alternate name forms together in useful concept clusters (Pyle 2016). More robust relationship-building is then necessary.

The GNA's Global Name Usage Bank is just such a mechanism. GNUB takes each individual name token reference and associates it with its full, institutional context. As explained by Pyle, the GNUB provides a persistent globally unique identifier (GUID) to all potential entities involved with the names instance, which includes agents (people and organizations that are responsible for the usage), as well as references, including published literature as well as unpublished reports, manuscripts, specimen labels, herbarium sheets, field notes, and the like, used in a taxon concept's circumscription (2016, 270–271). Within this network, any one reference document can contain many taxon name usages (TNU), just as any author can be associated with many published documents. With these associations in place, the disambiguation of one name-form from another can now proceed, which, given the size of the GNA, can be completed only through computational mediation.

Resolution

Linking harvested names to their sources, authors, and instances is just the first of many data transformations necessary to make this nomenclature useful. The next step, which emphasizes the syntactical elements of names, requires that scientific names be identified, extracted, and validated from all other nonstandardized name tokens. Whether or not colloquial or common names are integrated into the system must be decided at this point. Although the Catalogue of Life prioritizes code-compliant scientific names, common names are also widely available in its listing (see box 4.1). In a

Box 4.1

The Catalogue of Life has defined **fourteen field groups to be the standard set of data** for each species (or infraspecific taxa).

1. **Accepted Scientific Name** linked to **Reference(s)** (obligatory)
2. **Synonym(s)** linked to **Reference(s)** (obligatory, where available)
3. **Common Name(s)** linked to **Reference(s)** (obligatory, where available)
4. **Classification above genus, and up to the highest taxon in the database** (obligatory, where available)
5. **Distribution** (obligatory, where available)
6. **Life zone** (obligatory, where available)
7. **Current and Past Existence** (obligatory, where available)
8. **Additional Data** (optional)
9. **Latest taxonomic scrutiny** (obligatory)
10. **Reference(s)** (obligatory, where available)
11. **Taxon Globally Unique Identifier** (obligatory, where available)
12. **Name Globally Unique Identifier** (obligatory, where available)
13. **Catalogue of Life LSID** (obligatory)
14. **Source Database** (obligatory)

Catalogue of Life Standard Dataset Field Groups. Species 2000 has defined fourteen field groups to be the standard set of data (version 7, September 23, 2014) for each species and infraspecific taxon in the Catalogue of Life. (Species 2000 2016c). Used by permission.

perfect world, each name string is handled by human hands, though such attention is not possible in a pool of seventeen-million-plus names. Name matching services based on various lexical algorithmic software become crucial to facilitating this process (Pilsk, Kalfatovic, and Richard 2016; Vanden Berghe et al. 2015). The GNA has its own Global Names Resolver (GNA 2021b) that examines text strings to assess whether a name is in scientific form, correctly spelled, and currently in use, along with a host of other metrics (2020a). But algorithms and name resolvers have their limitations and can introduce any number of data failures and aberrations.

Scientific names, for example, must be Latinized and are most often binomial, consisting of two parts: a generic epitaph (genus name) and a specific epitaph (species name) (though variations to this rule do exist); *Ursus arctos* is an example of a correctly formatted scientific name. But as the GNA documentation points out, problems quickly arise during this process, since not all binomial Latinized name forms refer to species, such as with the terms anorexia nervosa and habeas corpus (GNA 2020b)—neither of which point to species, though either could be chosen by an algorithm as satisfying the syntactic standards for inclusion into a formal listing. Additionally, in some cases, orthographic conventions between the botanical and zoological codes sometimes call for different (and conflicting) nameforms that just cannot be resolved by automated means. Such shortcomings of GNUB are too significant for organizations such as the Catalogue of Life, and so alternate control mechanisms have been created to mediate these possible introduced errors.

The semantic value of names must next be addressed. Nomenclators identify the valid and accepted names that can be used for subsequent taxonomic work. But tracing the history and development of valid scientific names, and then subsequently disambiguating taxon concepts from one another is necessarily very human, time-intensive work. Priority requires understanding publication dates. To decipher one taxon concept over another, for example, nomenclators must examine all associated taxon documentation, including potentially the publication (with the detailed circumscription), as well as the type specimen. Nomenclator listings, then, are expected to be finely curated by way of expert review by assuring name currency and consistency and associating valid names with their associated publication dates (Croft et al. 1999, 320). Like thesauri, nomenclators are expected to establish primary taxon names and their synonyms. The result of this process is a controlled list of names that constitute code-governed facts. An example of a robust nomenclator is our previously discussed International Plant Names Index, located at the Royal Botanical Gardens, Kew. What emerges are clusters of related terms that contain a central, validated current nameform, as well as its associated synonyms and homonyms, similar to Pyle's final fish population, reproduced here:

Fish 1 and Fish 3: *Centropyge fisheri* (Snyder 1904) [sec.] Pyle 2003
> Fish 1: *Holocanthus fisheri* (Snyder 1904) [sec.] Snyder 1904
> Fish 1: *Xiphypops fisheri* (Snyder 1904) [sec.] Jordan 1922
> Fish 1: *Centropyge fisheri* (Snyder 1904) [sec.] Fraser-Brunner 1933
> Fish 3: *Centropyge flavicauda* (Fraser-Brunner 1933) [sec.] Fraser-Brunner 1933
= Fish 1: *Centropyge fisheri* (Snyder 1904) [sec.] Fraser-Brunner 1933
+Fish 3: *Centropyge flavicauda* (Fraser-Brunner 1933) [sec.] Fraser-Brunner 1933

Mapping out these name transformations over time makes for an exploitable information system in that users can assess, to a high degree of accuracy, the best species concept for their purposes. This synonymic amplification is a vital feature of effective biodiversity information seeking (Guala 2016).

THE CATALOGUE OF LIFE PLUS

To those designing the Catalogue of Life, it was clear early on that its hypercurated nomenclatural environment could be a hindrance to its widespread adoption and use. One can present many circumstances in which a perfectly valid taxon concept might not be present in the Catalogue. Recent publications describing new taxon concepts, for example, would not be represented in the system, given its update cycle. Further, there might not be enough human labor or expertise involved to meet increasing demand for more comprehensive lists. Still more onerous is to produce a list that includes a widely expanding groups of name variants, including a multitude of common names. The Catalogue of Life Plus is attempting to meet this challenge. This initiative is a partnership between Species 2000 and GBIF that resulted from the second gathering of the Alliance for Biodiversity Knowledge meeting (GBIF 2019). The CoL+ is meant to bridge an undifferentiated GNA-type space with the highly curated space of the Catalogue (Species 2000 2020 [2017]). The goals for the Catalogue of Life Plus are to

1. create both an extended and a strictly scrutinized taxonomic catalogue to replace the current GBIF Backbone Taxonomy and Catalogue of Life;

2. separate nomenclature (facts) and taxonomy (opinion) with different identifiers and authorities for names and taxa for better reuse;
3. provide (infrastructural) support to the completion and strengthening of taxonomic and nomenclature content authorities;
4. ensure a sustainable, robust, and more dynamic IT infrastructure for maintaining the Catalogue of Life.

The CoL+ is thus a layered clearinghouse system, whereby each layer represents a different level of curatorial control (see figure 4.2). The operational taxonomic units (OTU) in the outer cloud of figure 4.2 represent undifferentiated species concept tokens—a scientific name, a common name, or a genetic barcode. Users throughout the globe can then theoretically (given future infrastructural enhancements) propose relationships between OTUs in the outer cloud with scientific name-forms held in the middle cloud, here titled, "Col Plus/Linnean names." For example, an expert user in the genus *Abies* (a group including fir trees), could propose links between the

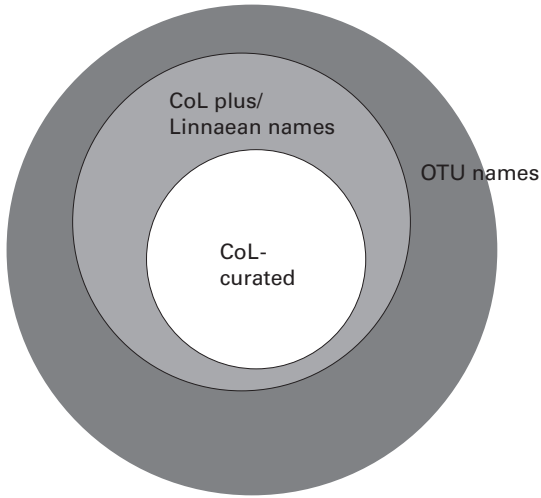


Figure 4.2
Catalogue of Life Plus Layer general schematic. Central names are curated to a gold standard. Linnaean names are names that are semantically meaningful and ready for ingest by the central repository. The outer ring consists of semantically incongruent forms that must be reconciled by volunteers before being ingested into the workflow.

common name "balsam fir" with its genetic barcode. Even more enticing, the CoL+ also grants the possibility for some a limited epistemic expansion of the system. What if, for example, individuals knowledgeable of indigenous botanical associations could also associate *Albies* with the Cherokee name for the same, *a'ninandak'*? Of course, this does not sidestep the general problematics of Western biological naming tradition (as will be discussed in future chapters), but it certainly does allow for more radical and diverse cultural enhancements of biodiversity knowledge. All of these associations can then be tentatively associated with the verified, scientific name *Abies balsamea* (L.) Mill in the CoL+ database layer. Such suggestions could then be flagged for validation by Catalogue editors and, after evaluation, allowed entry into the Catalogue's validated-name space. The result is a more nuanced, crowdsourced database that offer infinite opportunities to expand the horizons of nomenclature spaces.

In December 2020, the Catalogue of Life, in conjunction with GBIF, published a rebuilt infrastructure for what is now termed the Catalogue of Life Portal, which sets the stage for much of the work originally envisioned in the CoL+ (Huijbers 2020). The portal provides significant enhancements that situate the Catalogue for future growth and citation stability, particularly with respect to invalid names, with the prospect of slowly integrating these tokens into the core database set. A new ChecklistBank provides users the opportunity to upload databases into the Catalogue system, and thus an opportunity to publish this work to the broader taxonomic community (*CatalogueOfLife/ChecklistBank* 2021 [2017]). As part of the process of uploading subsidiary taxonomies, the data is normalized in accordance with Catalogue of Life's data model (Species 2000 2021a). All aggregated data is then available to the broader community through an application programming interface (API) to facilitate easy transfer of data to multiple database sources. Of technical note is that each species and higher taxa entry in the Catalogue database now has a persistent identifier that connects the current and archived data sets. With persistent identifiers in place, users and subsidiary databases can integrate the Catalogue and seamlessly update their repositories with each annual release without any identification conflicts. This new portal and ChecklistBank sets the stage

for a significant broadening of the core Catalogue nomenclatural set and taxonomic core.

The importance of developing systems such as the Catalogue of Life Plus cannot be underestimated, particularly because it is through these mechanisms that more inclusivity is built into the process of instantiative control. Concept control begins with the basic process of identifying the entities that constitute our field of concern, which, herein, I've operationalized as a process that involves producing species taxon concepts by way of the formal integration of evidence, producing nomenclatural tokens that are then appended to particular taxon circumscriptions, and subsequently mapping these name and taxon concept associations as scientific opinion evolves over a broad period of time.

The reality, of course, is that names make taxon concepts into, as Ron Day calls them, meaningful things, within a larger network of biodiversity practice (2014, 6). Names are tangible and machine-readable and, as tokens, are what systems are designed to manage and collocate (Furner 2016, 120), even if, as we will find, they fall short in many respects as monikers for complicated concept histories. I cannot stress the importance of this fact enough: our ability to accurately classify and represent the world (natural or otherwise) is directly proportional to the control of language terms that we have at our disposal to represent it. In chapter 5, we turn our attention to classifications-as-systems, including how they function as mechanisms of reduction, universality, and, ultimately, spaces of epistemic authority. As we have seen, the Catalogue is not only a space for nomenclatural control, it is also a taxonomic management system. It is here, at the level of the taxonomic whole, that we begin to see the structural expression of power take shape and the vast implications this has on scientific activity and the human imagination, broadly construed.

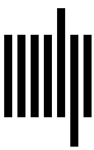
Power of Position

Classification and the Biodiversity Sciences

By: Robert D. Montoya

Citation:

Power of Position: Classification and the Biodiversity Sciences
By: Robert D. Montoya
DOI: 10.7551/mitpress/12245.001.0001
ISBN (electronic): 9780262369961
Publisher: The MIT Press
Published: 2022



The MIT Press

This book is freely available in an open access edition thanks to TOME (Toward an Open Monograph Ecosystem)—a collaboration of the Association of American Universities, the Association of University Presses, and the Association of Research Libraries—and the generous support of Arcadia, a charitable fund of Lisbet Rausing and Peter Baldwin, and the UCLA Library. Learn more at the TOME website, available at: openmonographs.org.

The MIT Press would like to thank the anonymous peer reviewers who provided comments on drafts of this book. The generous work of academic experts is essential for establishing the authority and quality of our publications. We acknowledge with gratitude the contributions of these otherwise uncredited readers.

This book was set in Adobe Garamond Pro by Westchester Publishing Services.

Library of Congress Cataloging-in-Publication Data

Names: Montoya, Robert D., author.
Title: Power of position : classification and the biodiversity sciences / Robert D. Montoya.
Description: Cambridge, Massachusetts : The MIT Press, [2022] | Series: History and foundations of information science | Includes bibliographical references and index.
Identifiers: LCCN 2021033972 | ISBN 9780262045278 (paperback)
Subjects: LCSH: Biology—Classification. | Life sciences—Classification. | Cladistic analysis.
Classification: LCC QH83 .M68 2022 | DDC 570.1/2—dc23/eng/20211221
LC record available at https://lccn.loc.gov/2021033972