OMICS A Journal of Integrative Biology
Volume 7, Number 1, 2003
© Mary Ann Liebert, Inc.

Future-Proofing Biological Nomenclature

GEORGE M. GARRITY[1] and CATHERINE LYONS[2]

TO MOST BIOLOGISTS, it seems inconceivable that the simple act of naming a biological entity has any more significance than identifying a personal achievement or staking a claim to a territory of research interest, akin to carving one's initials into the tree of life. However, this simple act has potentially far-reaching and long-lived consequences. Names, especially those ascribed to organisms, serve as a primary entry point into the scientific, medical, and technical literature and figure prominently in countless laws and regulations governing various aspects of commerce, public safety and public health. These names also serve as a primary entry point into many of the central databases that the scientific community and the general public now rely upon.

Even well-formed and properly applied names can serve as a source of confusion and considerable frustration. This is hardly a new problem. In the mid 18th century, Carl von Linné proposed the use of latinized binomial names and a hierarchical classification scheme as an alternative to trivial and colloquial names, which were a constant source of confusion among his contemporaries (Linné, 1735). The Linnaean system of nomenclature was widely accepted and has subsequently been codified into four separate legalistic frameworks (the codes of botanical [Greuter, 2000], zoological [Ride, 1985], prokaryotic [Sneath, 1992], and viral nomenclature [Buchen-Osmond, 2002]), describing the rules for forming and ascribing names to species and higher taxa, circumscription and emendation, priority and citation, synonymy and more recently homonymy, correction of orthographic errors and adjudication of disputes in nomenclature. Chief among the stated objectives of these codes is stability and order in the nomenclature of the taxa covered by each. Achieving this goal, however, remains elusive. There are a variety of reasons.

Relatively few contemporary biologists are actively engaged in systematic studies and have reason to formally propose names for new biological entities. Rather, they are end-users of the classifications and nomenclature produced by a small group of specialists. Most biologists fail to recognize that taxonomic proposals are expert opinions, arising from comparative studies of small numbers of species that may or may not be representative of larger natural groups and that the opinions rendered by these experts are subject to acceptance or rejection by the larger community of biologists. They also fail to recognize that the Codes of Nomenclature do not govern the process of biological classification or identification, only the formation and assignment of names to proposed taxa. Legitimate and valid names may be ascribed to poorly formed taxa and illegitimate and invalid names may be assigned to well formed taxa.

The name ascribed to a given group is fixed in both time and scope and may or may not be revised when new information is available. When taxonomic revisions do occur, resulting in the division or joining of previously described taxa, authors frequently fail to address synonymies or formally emend the descriptions of higher taxa that are affected.

Whereas the different codes of nomenclature guarantee persistence of a formal name, the serial, cumulative nature of effective and valid publication allows the name to obsolesce in relation to the taxon it originally denoted. In contrast, it is the taxon itself that persists, and the granularity with which it is defined increases over time. The formal name provides an archival record of taxonomic definition only for a single point in time—the date of publication. A robust and persistent taxonomy requires taxonomic definition to

[1]Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, Michigan.
[2]Explicatrix, Stirling, New Jersey.

be a maintained, networked resource, rather than a retrospective sequence of names and emendations. A commonly referenced terminology based on persistent, increasingly refined taxa is needed to replace or augment a static nomenclature that diverges over time from the taxonomy it initially denotes.

This disjunction of nomenclature and taxonomy results in an accumulation of names of dubious value in the literature and databases. While systematic biologists may be adept at recognizing such problems, most others (including the curators of some databases) are not. This can have a significant impact on activities such as assertions of taxonomic identity, commonality of metabolic function, and recognition of homologous, paralogous or xenologous genes. It can also have significant and unintended consequences such as adding or removing species to lists of tightly regulated species (e.g., the current list of biothreat agents).

There are a number of significant distinctions between the more mature Botanical and Zoological Codes of Nomenclature and the Code of Prokaryotic Nomenclature. The record of botanical and zoological nomenclature dates back to the time of von Linné, and names are considered valid regardless of where they were published, provided that other criteria are fulfilled. Each worker is obliged to establish the priority of names that are used in their proposals. While lists of names in common usage are available as a starting point (Melville and Smith, 1987; Buchen-Osmond, 2002), the number of synonyms for a given group can easily exceed the number of names deemed legitimate. Prokaryotic biologists have addressed this problem in a much more elegant manner. In 1980, those names that were considered of dubious value were purged from the record, and the Approved Lists of Bacterial Names were published (Skerman et al., 1980). These officially sanctioned names represented a new starting point, and a formal mechanism for adding new names to the record was established; in essence, a registry was created. The effects of this mechanism prove instructive in other ways, as well.

Since publication of the Approved Lists, the number of validly published names of species and higher taxa has increased dramatically. This can be attributed to the widespread application of high-resolution molecular methods to resolve taxonomic problems and to infer evolutionary relationships among *Bacteria* and *Archaea.* While these studies have led to the first natural taxonomy of prokaryotes (Garrity and Holt, 2001; Garrity et al., 2002), a side effect has been a degradation of nomenclature. By the end of 2002, one-fifth of the more than 6,500 validly published species names were reduced to synonymy. In the absence of an *a priori* knowledge of the biological nomenclature, it becomes increasingly unlikely that even the most sophisticated users will be able to perform a complete and accurate search of either the scientific, technical or regulatory literature or databases. There is a clear and unfulfilled need for a mechanism that integrates this critical information into a networked, distributed environment to circumvent this problem.

As biological data proliferates and interconnects, it depends increasingly on software infrastructure, and it becomes increasingly obvious that biological names do not meet the requirements of a good identifier, in strict computing terms. A good identifier should be unique and persistent. Ideally, each name should point to a single time-specific or publication-specific definition of a taxon (although if necessary each taxon can have more than one identifier). In reality, only the lower taxonomic entities are persistent and permanently coherent, phylogenetically. As new data become available, the inferred relationships among the named entities may change: a taxon may be promoted or demoted, new taxa may be interposed between formerly contiguous taxa. While such events should trigger a full and automatic updating of the definitions of the affected names at all levels, that rarely occurs. As a result, the formal association of names with taxa tends to weaken as the rate at which gene sequencing accelerates. (Gene sequencing is increasingly used to define and delineate taxa.) Failure to address this problem will result in increasingly unpredictable responses when biological names are used to query either the literature or databases. What is required is a resolution system that can handle the complex relationships between biological names and the entities they denote and provide links to both the historical and current definition of each named taxon.

We believe that an implementation of the Digital Object Identifier (DOI, [Paskin, 2002; Paskin, 2003]) may provide the most robust and future-proof solution to this problem. A DOI is a unique, persistent identifier of an information resource that is registered together with a URL. Its purpose is the management and retrieval of that resource in the networked environment. In practice, most current DOIs identify journal articles, but DOIs are now being applied to trade publications, stock photography, and physicochemical data sets.

We are developing a model for assigning DOIs to prokaryotic taxa as a test case. Though the definition

of a taxon may be refined and its nomenclature redefined, the DOI will persist, leaving a forward-pointing trail that can be used to reliably locate digital and physical resources, even when a name may be deemed obsolete. Forward linking from a synonym to a record of the publication that asserts synonymy is especially important, as there is currently no mandatory mechanism for asserting and resolving names that become ambiguous. Our model seeks to strengthen the association of names with taxa by using DOIs to track the taxonomic definition of a name over time. It is extensible to the level of individual genes within a given species. However, the real power of this method lies in the ability of DOIs to become embedded in the information environment, providing a direct and persistent link to the full record of taxonomic and nomenclatural revision and ensuring consistency and accuracy throughout online scientific resources. In building a DOI-based infrastructure for formally associating nomenclature with taxonomy, we envision a time when a name can be used unambiguously and persistently, only one mouse-click away from a record of its current definition and historical development.

REFERENCES

BECKER, E., GUNNEWIG, D., BUHSE, W., et al. eds. (2003). *Digital Rights Management: Technical, Economical, Judicial, and Political Aspects.* (Springer-Verlag, Heidelberg).

BUCHEN-OSMOND, C. (2002). ICTVdB: The Authorized Universal Virus Database, Biosphere 2 Center, Columbia University. Available: http://ictvdb.bio2.edu/index.htm.

GARRITY, G.M. and HOLT, J.G., eds. (2001). *The Road Map to the Manual. Bergey's Manual of Systematic Bacteriology.* (New York: Springer-Verlag).

GARRITY, G.M., JOHNSON, K.L. et al. (2002). *Taxonomic Outline of the Prokaryotes, Bergey's Manual Of Systematic Bacteriology, Bergey's Manual Trust.* DOI 10.1007/bergeysoutline.

GREUTER, W. (2000). *International Code of Botanical Nomenclature (St. Louis Code).* (Koeltz Scientific Books, Koenigstein, Germany).

LINNÉ, C. v. (1735). *Regna Tria Naturae Systematice Proposita per Classes, Ordines, Genera, & Species.* (Stockholm: Lugduni Batavorum).

MELVILLE, R.V., and SMITH, J.D.D., eds. (1987). *Official Lists and Indexes of Names and Works in Zoology.* (London: International Trust for Zoological Nomenclature).

PASKIN, N. (2002). *Digital Object Identifiers. ICSTI Seminar: Digital Preservation of the Record of Science.* (IOS Press, Amsterdam).

PASKIN, N. (2003). DRM technologies: identification and metadata. *Lecture Notes in Computer Science. Digital Rights Management: Technical, Economical, Judicial, and Political Aspects.* E. Becher, D. Gunnewig, W. Buhse, and N. Rump, eds. (Springer-Verlag, Heidelberg).

RIDE, W.D.L., ed. (1985). *Code International de Nomenclature Zoologique.* (Berkeley: University of California Press).

SKERMAN, V.D.B., McGOWAN, V., et al. (1980). Approved lists of bacterial names. Int. J System. Bacteriol. **30,** 225–420.

SNEATH, P.H.A., ed. (1992). *International Code of Nomenclature of Bacteria (1990 Revision).* (Washington, DC: American Society for Microbiology).

Address reprint requests to:
Dr. George M. Garrity
Department of Microbiology and Molecular Genetics
Michigan State University
6162 Biomedical Physical Sciences Building
East Lansing, MI 48824-4320

E-mail: garrity@msu.edu