Invited Essay

Syst. Biol. 55(3):367–373, 2006
Copyright © Society of Systematic Biologists
ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150500541680

Taxonomic Indexing—Extending the Role of Taxonomy

DAVID J. PATTERSON, DAVID REMSEN, WILLIAM A. MARINO, AND CATHY NORTON

Marine Biological Laboratory, Woods Hole, Massachusetts 02543, USA; E-mail: dpatterson@mbl.edu

Abstract.— Taxonomic indexing refers to a new array of taxonomically intelligent network services that use nomenclatural principles and elements of expert taxonomic knowledge to manage information about organisms. Taxonomic indexing was introduced to help manage the increasing amounts of digital information about biology. It has been designed to form a near basal layer in a layered cyberinfrastructure that deals with biological information. Taxonomic Indexing accommodates the special problems of using names of organisms to index biological material. It link alternative names for the same entity (reconciliation), and distinguishes between uses of the same name for different entities (disambiguation), and names are placed within an indefinite number of hierarchical schemes. In order to access all information on all organisms, Taxonomic indexing must be able to call on a registry of all names in all forms for all organisms. NameBank has been developed to meet that need. Taxonomic indexing is an area of informatics that overlaps with taxonomy, is dependent on the expert input of taxonomists, and reveals the relevance of the discipline to a wide audience. [Biodiversity informatics; names; taxonomic indexing; taxonomy.]

INTRODUCTION TO TAXONOMIC INDEXING

Taxonomic indexing is a new area of biological informatics. It is one of a number of taxonomically intelligent network services that use nomenclatural principles and elements of expert taxonomic knowledge to manage information about organisms. Taxonomic indexing addresses the need to manage rapidly growing amounts of information about organisms in a knowledge environment that is increasingly digital and heterogeneous (Agosti and Johnson, 2002; Patterson, 2003; Stein, 2002).

Taxonomic indexing services treat the names of organisms as 'metadata' terms capable of defining subsets of information. One or more of these terms are associated with all items of information about taxa such as scientific papers and museum or herbarium records. When the names are recorded, indexed, and properly managed, they can be used to retrieve and organize the source records (these are collectively known as "name-bearing data objects"). Collectively, all names provide comprehensive metadata coverage for all information about all named taxa.

Simple name-based indexing systems use only the names that are found in the target array of name-bearing data objects—rather like the index of a book. Taxonomic indexing enhances simple name-based indexing in a variety of ways. For example, taxonomic indexing acknowledges that there may be many different names for the same taxon and that these need to be linked so that a query starting with one name will find information tagged with other names. This enhancement is referred to as "reconciliation." Secondly, taxonomic indexing recognizes that the same name may be used for more than one taxon and that the indexing service must resolve the resulting ambiguity so as not to merge data on different entities (this is "disambiguation"). Thirdly, indexing services need to represent the factual association of one name with one or more data objects. It can do this by extracting and separating the names from the data objects and then cataloging the relationship among names and data. If taxonomic indexing is to index all information, it must also recognize vernacular names and misspellings as metadata because it must have access to a list of all names that have been used for all organisms.

THE NEED FOR TAXONOMIC INDEXING

The taxonomic impediment describes the inadequacies of the current taxonomic infrastructure (Environment Australia, 1998). The most common metric by which the taxonomic impediment is measured is the number of professionals who regard themselves as taxonomists (Hopkins and Freckleton, 2002). This community has been estimated as having about 6000 members (Wilson, 2003). The ETI taxonomist database has about 4200 registrants (http://www.eti.uva.nl/tools/wtd.php). This is well under half the number necessary to provide minimal coverage for all biodiversity (Hebert et al., 2003).

Despite the decline in the number of taxonomists, the rate of species discovery remains unchanged (e.g., http://data.acnatsci.org/wasp/; Froese and Capuli, 2005; Knapp et al., 2005; Saarenmaa, 2002; Wilson, 2003). This is attributable to improvements in species discovery made possible by molecular technologies (Hebert et al., 2003; Ventner et al., 2004).

These conflicting metrics reveal the taxonomic impediment to be multifaceted. An aspect of taxonomic

infrastructure that is not enhanced by molecular technologies is the custody and management of biodiversity information (Lee, 2000; Wheeler, 2004). The amount of biodata is vast. Each of the 2,000,000 or so described species expresses a biology that extends from biochemical pathways to their role in ecosystems. Sequence data housed in databases now amount to 10^{11} to 10^{12} bases and are growing exponentially. Ecological and distributional data are being generated in ever-increasing amounts from telemetric data and satellite monitoring. New molecular tools can generate millions of reports of the occurrence of organisms in a matter of hours (Margulies et al., 2005). Massive bodies of data traditionally documented in the form of books and papers are being made available through digitization initiatives by (among others) J-STOR, Google, GBIF, and the Sloan Foundation. Yet, the proportion of this information that is available through the Web in an integrated fashion is paltry (Godfray, 2002; Saarenmaa, 2002). If we are to gain full benefit from the availability of this information, we need to improve our management of biological inforination. Taxonomic thinking has managed information very successfully from the time of Linnaeus. We have sought to embed nomenclatural and taxonomic principles in new tools to manage information about organisms.

A Layered Architecture of Knowledge

Our goal in information management should be to reconstitute biological knowledge by emulating the relationships among individual data elements. We find it simplest to conceive of knowledge and associated services within a "layered" structure (Fig. 1). This structure recognizes that knowledge is comprised of factual (objective or universally agreed) and subjective (views) elements. In a layered environment, factual information is segregated and placed in a common and basal pool. As there is no dispute over facts, basal layers are well suited to being assembled and maintained by the community at large. This agreed pool of information can be called upon by many different users and used for different purposes. Subsequent layers add value to the factual information by selecting facts, combining, and annotating them. The separation of factual content removes the need to duplicate repositories of factual information. The dependency of the end users on the veracity of underlying data exposes the content to scrutiny, and ensures high quality standards.

Treating Names as Metadata

"All accumulated information of a species is tied to a scientific name, a name that serves as the link between what has been learned in the past and what we today add to the body of knowledge." (Grimaldi and Engel, 2005). Each item of information is found in what we refer to as 'name-bearing data objects.' Name-bearing data objects are recorded items in, for example, the primary taxonomic literature, specimen collections, floras, faunas, ecological documents, molecular data bases, aboriginal

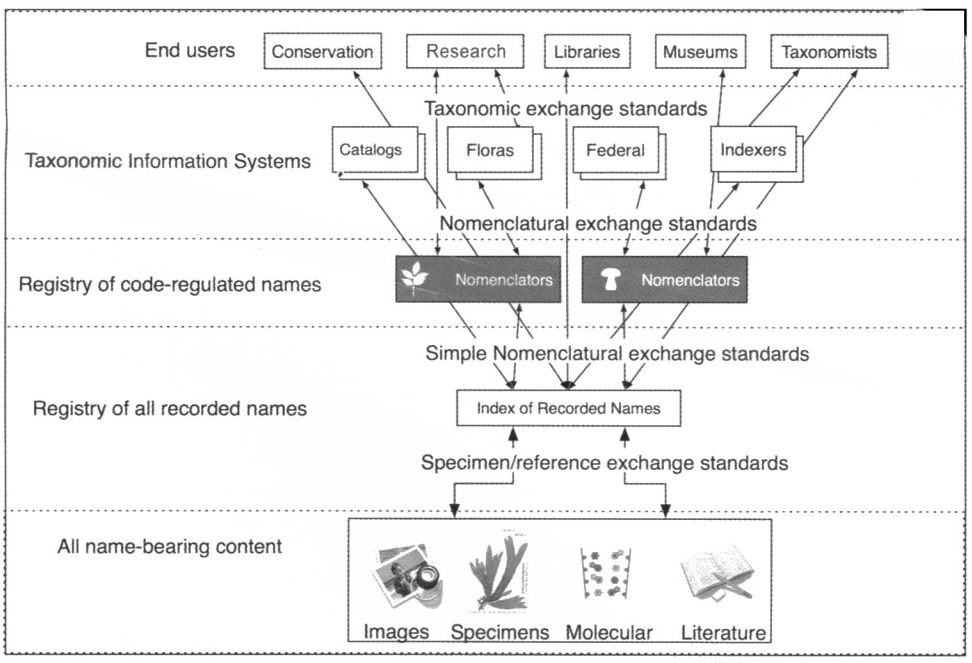

FIGURE 1. Information management within a layered architecture. The basal layer includes all objects that carry information about organisms such as items in biological collections or publications. Collectively, these objects contain all known names for organisms, and these can be segregated into a second factual layer—a registry of names. Higher layers in the architecture add value to this compilation—such as selecting those elements that comply with the codes of nomenclature. This information is further enhanced by biodiversity information specialists—such as producers of catalogs, floras, faunas, and other biota lists; governmental agencies and biological information services. Those activities generate enhanced products that serve many categories of end users.

knowledge structures, and so on. A name, such as *Xysticus cristatus*, within or associated with a data object that exists in a recorded form, can be exploited as a label for the information and the object. The same name can be used to label other objects. In that way, names of organisms are a category of metadata capable of organizing information about organisms in all name-bearing data objects. Indexing services that intend to work with all information that is already cataloged or has yet to be cataloged will require access to all of the names that have been used for all organisms.

The uBio project (http://www.ubio.org) is an initiative to gather all names for indexing purposes. To estimate the total number of names, we note that flowering plant databases, which are probably the most comprehensive databases with names, hold about five times more names than there are species (http://www.ipni.org/ik_blurb.html, http://www.tropicos.org/). From this, we estimated that there are about 10,000,000 code-compliant names out there. This number does not remain static as about 1% of all formal names change each year and there is a further 1% expansion through new discoveries (Biosis, 2005; Froese and Capuli, 2005). This number is still less than 10% of the names that are required for indexing purposes. For indexing purposes we need not only correct code compliant names, but their misspellings, lexical variants (Table 1), and other alphanumeric labels such as sample and culture identifiers. Also required are vernacular names in any of more than a thousand languages and written in one of 150 or more alphabets, machine-generated errors such as truncations and other aberrations. All have value as metadata and all need to be collected and curated because they will be needed to find data objects in the e-world.

General aggregators of code-compliant names, such as Species 2000, ITIS (collectively forming the Catalogue of Life; Bisby et al., 2005), and GBIF with the Electronic Catalogue of Names of Known Organisms, have progressed beyond the milestone of 1,000,000 names and half a million species. These lists include or are supplemented with on-line compendia of most genera of viruses (Büchen-Osmond, 2005), prokaryotes named in compliance with the current code of nomenclature (Euzéby, 2005), protists (http://microscope.mbl.edu), algae (Guiry and Nic Dhonncha, 2005), plants (Farr and Zijlstra, 2005; http://www.ipni.org/ik_blurb.html), fungi (www.indexfungorum.org), and, with the recent posting of Nomenclator Zoologicus (Neave, 1939–1996) online (http://uio.mbl.edu/NomenclatorZoologicus/), a fair proportion of animal genera. Many other enterprises such as Index Fungorum, IPNI, AntBase (http://www.antbase.org/), and the Diptera site (Thompson, 2005) provide species-level coverage for selected taxa. Compendia of geographically or ecologically filtered information such as the Australian ABIF site (http://www.deh.gov.au/biodiversity/digir/), European marine species (ERMS; http://erms.biol.soton.ac.uk/), North American insects (http://www.nearctica.com/nomina/main.htm), organisms from New Zealand (Manaaki Whenua Landcare Research, 2005) or Costa Rica (http://darnis.inbio.ac.cr) also fill in species-level details. Agencies such as museums, herbaria, bibliographic services, and molecular databases all independently compile names of locally held assets.

The growth of these lists is slow, largely because the process includes an early and time-consuming step of vetting the name by taxonomic criteria (Fig. 2; Patterson, 2003). As this is not necessary for indexing purposes, we accelerated the rate of names acquisition by removing this bottleneck. Vetting is deferred to a later stage. We assigned priority to generic names because the binomial character of species names ensures that generic names are sufficient as metadata for all scientific information on taxa. The use of generic names gives an order of magnitude less resolution than species names, but a list of names of genera is at least two orders of magnitude easier to assemble. We refer to our approach

TABLE 1. Lexical variants of names. Some lexical variants of names of the false foxglove in the International Plant Name Index (http://www.ipni.org/), United States Department of Agriculture Plants Database (http://plants.usda.gov/), Robert W. Freckmann Herbarium at the University of Wisconsin (http://wisplants.uwsp.edu/), and the Ohio Department of Natural Resources (http://www.dnr.state.oh.us/).

Variant form of name	Source
Gerardia paupercula var. *borealis* (Pennell) Deam	IPNI
Gerardia paupercula (Gray) Britt. var. *borealis* (Pennell)	USDA
Gerardia paupercula (A. Gray) Britton var. *borealis* (Pennell) Deam	OHIO DNR
Gerardia paupercula (A.Gray) Britton subsp. *borealis* (Pennell) Pennell	Freckmann
Gerardia paupercula Britton subsp. *borealis* Pennell	IPNI

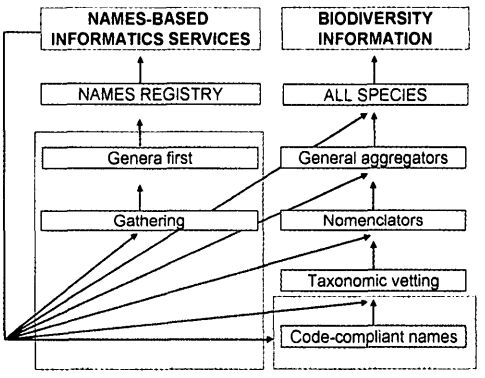

FIGURE 2. Filling the pool of names. Two strategies are illustrated. The traditional or taxonomic approach is to the right. Names are extracted from name-bearing data objects. Code-compliant names are selected and subject to critical scrutiny to generate reviews and nomenclators. The correct names and their synonyms find their way to aggregators. The objective is a list of all species, and this serves the information needs of biodiversity specialists. The alternative approach to the left (gathering) emphasizes the acquisition of names, defers the taxonomic input, gathers generic names first, and uses the resulting pool of names to underpin a diverse array of biodiversity informatics services. Some name-based services, such as automated names discovery and mapping of names, enhance the gathering and taxonomic activities. The heights of the boxes at the bottom reveal the relative numbers of names assembled by exemplars of two approaches, uBio and Species2000.

as "gathering" and targeted global lists of genera that were available on the Internet and in the primary literature. We extracted names by a variety of ways, from retyping documents to working with taxon-specialists in harvesting and parsing lists. Automatic name-finding tools are also becoming available (Koning et al., 2005). Gathering created a compilation of most generic names for living organisms within a few years. This list has been converted into a specialist lexicon that is now used in conjunction with in-house automated names discovery tools (http://names.mbl.edu/tools/index.html). These tools (hunter gatherers) accelerate names acquisition by at least a further order of magnitude. The result is sufficient coverage (well over 5,000,000 records) for indexing tools that serve most biodata to come into operation now. With enthusiastic donations by all players who hold name information into the common pool, and through continued use of hunter gatherer tools, species-level resolution can be achieved for 95% of current biodiversity information within a short time span.

NameBank—The Registry of Biological Names

Once compiled, names need to be assigned to a database and an environment that is custom-designed to catalog and manage names and information about names. Such environments are called "name servers." For taxonomic indexing, a name server must be able to embrace any name in any form and it should be comprehensive and capable of holding all names for any organism. In our view, the data model should comply with the relevant metadata schemas (the Linnaean Core is the most relevant; Hobern, 2005). Name servers should segregate subjective and objective information (Pullan et al., 2000), be capable of adding value through services such as reconciliation and disambiguation, be placed within a Web-accessible environment capable of permitting community participation in the assembly and vetting of the names compilation, and operate in a way that tracks contributions to ensure that all participants are properly accredited. Most importantly, the name server must have a service mentality to serve any and all communities of users who have interests in names. These requirements led to the layered approach described above (Fig. 1).

The name server compilation of names that underpins taxonomic indexing is NameBank (http://www.ubio.org/nameserver/DataModel.htm). It will be described in a later publication. NameBank holds only objective information: names, their sources, and objectively defined relationships among names. It has over 5,000,000 records at the time of writing. It protects the indexing benefits of the more subjective elements of taxonomy, such as hierarchical relationships and the mapping of heterotypic synonyms, by placing this subjective information in an associated database called ClassificationBank where concepts may also be housed. Navigability among all names is achieved by unifying the names structure within an environment that can depict an indefinite number of differing classifications.

Because it can act as a common pool of agreed name information that may be shared among many clients, NameBank serves as a general and universal biological names registry. A universal biological names registry is inherently unifying, will promote machine-to-machine dialogue, can become the mechanism through which code-compliant names are introduced, and, when fully populated and annotated, can disambiguate homonyms (Thorne, 2003; Patterson et al., 2003). Because of its comprehensive and objective character, it does not need to be duplicated. The needs of different stakeholders can be achieved by applying additional layers atop of this registry. Such layers might, for example, extract subsets of names to provide more specialist services such as those of nomenclaturalists (Polaszek et al., 2005) or for compliance with the Phylocode (http://www.ohiou.edu/phylocode/).

Using Names for Indexing

Names are not perfect instruments for indexing. They may be ambiguous, they may lack full resolution, or they may be bundled up with opinions that undermine the objectiveness of the indexing service.

Setting aside issues of misidentification and erroneous data entry (Chavan et al., 2005), ambiguity arises for several reasons. First, there may be many alternative names for the same entity. Evolutionary insights may result in a species being moved from one genus to another (*Peranema fusiforme* becomes *Jenningsia fusiforme* as newer observations correct earlier errors based on knowledge gaps). This action creates homotypic or objective synonyms. Should someone hold the view that the names *Jenningsia fusiforme* and *Jenningsia macrostoma*, initially described as different species, refer to the same organism, these would be heterotypic or subjective synonyms. Both cases create situations where data on the same organism are labeled with different names.

The consequences of not using all alternative names is poor performance in indexing. A keyword search of the biomedical citation database PubMed for the newt, *Notophthalmus viridescens*, retrieves (at the time of writing) 350 citations, the earliest of which dates back to 1965. However, this newt has many names, some of which are indisputably linked to each other as homotypic (objective) synonyms, lexical variants, or misspellings (Table 2). Unless these names are also included in a

TABLE 2. Recovery of records using different objectively related names. Recovery of records from PubMed and JSTOR with 5 of the 15 known objectively related names that have been applied to the red spotted newt from eastern North America. The last two names are misspellings.

Name	Date that the name was first used	Items in PubMed	Items in JSTOR
Notophthalmus viridescens	1965	350	281
Diemictylus viridescens	1959	36	38
Triturus viridescens	1949	87	280
Diemyctilus viridescens	1965	1	3
Diemyctylus viridescens	1964	3	70

search, about one quarter of the items known to PubMed are not recovered and the proportion of unrecovered material increases to over half when indexing is extended to older source material that is available through JSTOR. This problem is overcome with taxonomically intelligent indexing, because homotypic synonyms, their lexical variants, and associated vernacular names are mapped together within a reconciliation group. A reconciliation group serves to convert a query initiated with one name for an entity into an action using all names that have been used for the entity. Reconciliation offers a mechanism to allow enterprises such as TreeBase that acquire data sets from many sources but that do not use the same labels for taxa (Herbert et al., 2004).

A second indexing problem arises with names that are spelled identically but refer to different taxa—homonyms. The name *Peranema* was introduced by Don to refer to a fern (Don, 1825) and by Dujardin to refer to a flagellate (Dujardin, 1841). The codes of nomenclature eliminate the duplicates by endorsing the first use of the name and requiring the later names to be replaced. Each code of nomenclature only deals with some organisms, and each (the current code for prokaryotes) is blind to names introduced under other codes. This situation permits the legitimate introduction of homonyms. Fixing the homonym problem can make the situation worse. In the case of *Peranema*, the flagellate is a euglenid, an ambiregnal territory in which nomenclature can legitimately be handled under more than one code (Patterson and Larsen 1991). The rules of the International Code of Botanical Nomenclature apply to this name and require the name of the euglenid, as the later homonym, to be changed (it was changed to *Pseudoperanema)*. The International Code of Zoological Nomenclature also applies, but does not recognize Don's name because it refers to a fern, and only considers the name of the euglenid. From that point of view, the ICZN requires no change to the name of the flagellate. The consequences are that the name *Peranema*, from the botanical perspective, legitimately refers to a fern, and, from the zoological perspecitive, to a protist. The euglenid genus has more than one legitimate name (*Peranema* and *Pseudoperanema*) and has more than one type species (*Peranema trichophorum* Dujardin, 1841 and *Pseudoperanema hyalinum* Christen 1962) (Larsen and Patterson, 1991).

Homonymy is a problem that is mostly expressed at the level of genus. About 13% of all botanical generic names are homonyms of zoological names (McNeill, 1997). Name-based services must resolve the ambiguity caused when the use of a homonym can draw together information about two types of organism. Disambiguation may be achieved by reference to broad taxonomic territories (*Peranema* Pteridophyta is not the same as *Peranema* Protista) or to families. Another means of disambiguating homonyms is to include the authority (*Peranema* Dons 1825 versus *Peranema* Dujardin 1841). This works in almost all cases, but in a few instances the same author has introduced the same name in the same year for different taxa within the same family.

The problem with homonyms is confounded by chresonyms. Chresonyms are references to the use of a name. They can be presented in many formats (*Jenningsia fusiforme pro parte* Patterson, 1995, or *Jenningsia fusiforme* in Patterson, 1995). Problems arise when the format is simply "Name user" (such as *Homo sapiens* Smith, 2005). This is intended to indicate Smith's use of *Homo sapiens* in an item published in 2005). This form is not distinguishable in form from code-compliant names. These chresonyms are often included in lists of synonyms. Vetted compilations of code-compliant names can contain numerous chresonyms (the 2005 version of the Catalog of Life has 52 entries for *Xysticus cristatus).* The occurrence of the same spelling for genus and species but with different authors is sufficient to alert us to the presence of chresonyms, and the need for taxonomic intelligence to disambiguate them from homonyms.

A fourth complication of using names for indexing is that there may be different views as to what names refer to ("taxonomic concepts"; Geoffroy and Berendsohn, 2003). As an example, the names *Jenningsia fusiforme* and *Jenningsia macrostoma* were initially used for what were believed to be different species. Should these be regarded as being indistinguishable, then the two names would refer to the same organism. The consequence would be that *Jenningsia fusiforme* refers to two sets of name-bearing objects, one excluding anything relating to *Jenningsia macrostoma*, and one including that content. Unless the concept is specified in some way, we do not know if the sentence "*Jenningsia fusiforme* has a worldwide distribution" excludes the entities described under the name *Jenningsia macrostoma*, or includes them. Our assessment of the literature is that much fewer than 5% of the name-bearing data objects specify a taxonomic concept.

Application of Taxonomic Indexing

Taxonomic indexing has the potential to interconnect all information about organisms that can be accessed through the Internet in a biologically meaningful way. We are exploring this potential in a variety of ways. STAR Web sites are biological content management systems that can be modularized and assembled into extensive networks capable of acting as a medium for the "Encyclopaedia of Life." These sites interconnect local and distributed data using a Taxonomic Indexing core comprised of the unified classification of names. The first star*site is micro*scope (http://microscope.mbl.edu). It is built with a layered architecture so that it can provide selective content to other sites such as Microbial Life (http://microbial.life.mbl.edu) and the International Census of Marine Microbes (http://icomm.mbl.edu). Other proof of concept devices include applications that enhance generic search engines such as Google (http://tns.mbl.edu/clients/google/index.php) or the JSTOR online journal service (http://uio.mbl.edu/clients/jstor/) with expert taxonomic knowledge. We have also demonstrated the utility of the system in developing a browsable version of the 5000-page Birds of

the Belgian Congo (http://www.ubio.org/services/amnh/amnh.html). This contains over 10,000 distinct names, one of which refers to a genus of fly (Diptera). A keyword search on the word "fly" or "flies" in a book relating to flying vertebrates is unlikely to discriminate the dipteran. It would be prohibitive to search for all quarter of a million dipteran names already cataloged (http://www.sel.barc.usda.gov/Diptera/biosys.htm). However, automated indexing tools that place names within a hierarchical superstructure can find the page with the single reference to a fly with three clicks of a mouse. We are also building automated names discovery tools, names processing tools, and tools, such as LinkIT that cross-links data sources (http://names.mbl.edu/tools/index.html), or a taxonomically intelligent feeder reader than indexes RSS feeds on the fly (http://www.ubio.org/index.php?pagename=ubioRSS), or that use aggregation technology to generate species pages from distributed data sources (http://portal.ubio.org/).

Copyright, Plagiarism, and Credit

The process of compiling names for taxonomic indexing raises questions about whether copyright protection applies to names. Copyright protects original creative expression. Names of organisms are factual elements and this precludes them, whether singly or in compilations, from being protected by copyright considerations. This position was broadly established in a dispute over telephone numbers and client addresses (http://www.justia.us/us/499/340/). From this it follows that the reassembly of names of organisms from any sources does not infringe any copyright.

There is a view that the "sweat of brow" required to create a compilation provides the compiler with intellectual rights. This leads to complaints of "plagiarism" when the factual content of compilations is used by someone other than the compiler without acknowledgement. The sweat of brow argument has a different status in different countries, but does not have legal standing in most (e.g., http://www.gesmer.com/publications/softcopy/15.php). Contemporary classifications are mostly comprised of hierarchical schemas that have been inherited from predecessors. These have been openly shared through the scientific literature and traditionally we acknowledge the contributions of individuals by citing them. This creates a credit trail that appropriately identifies the efforts of others. The credit trail can be protected within name-based services. To do this, the contributors of names to the registry are identified so that the user of their names by services is recorded and can be reported back to the supplier.

The Vision

Biodiversity informatics services will emerge for many end users in the very near future. Information professionals such as publishers and librarians will have access to automated indexing services. These and Internet search engines will reconcile names in real time and can exploit hierarchies to offer search enhancements that focus or generalize searches. Students and researchers will have access to more and better vetted information. General users and information managers will no longer need to remain current with the most recent nomenclatural changes. Agencies providing commercial services will find those services enhanced. Taxonomists will create more accessible and communally owned repositories of authoritative taxonomic information and opinion, and will move more rapidly to consensus classifications to facilitate the census of life. Layers on top of registries can form reference structures that will improve the quality and precision in taxonomy. By linking name-based services with tree-based thinking and phyloinformatics, phylogeny and molecular catalogs will be integrated into traditional knowledge. The emergence of name-based services will reduce the load carried by taxonomists and alleviate one dimension of the taxonomic impediment.

The emergence of a comprehensive names registry will create a significant change in the taxonomic landscape. Taxonomic indexing and other name-based services are in their infancy and just beginning to emerge from a small number of innovators, but their utility makes widespread adoption inevitable. It is not clear who will own this domain. Unlike taxonomy, the registry and associated name-based services will not be partisan—that is, they will not be restricted to subsets of life. A leadership role for taxonomists is NOT inevitable. To retain control of this field, taxonomists will need to welcome and manage all kinds of names, misspellings, and alphanumerical identifiers and treat them as equal in value for indexing purposes. New tasks will include the assembly of reconciliation groups, the expert disambiguation of homonyms, and the addition of layers that annotate names of relevance to traditional taxonomy, or that segregate chresonyms from synonyms. The new beneficiaries of name-based informatics will include commercial information managers and this should lead to alliances between taxonomists and for-profit organizations that can offer a new income-line to the custodians of the discipline. If the responsibility for the development and application of taxonomically informed tools is not retained by taxonomists, it will lead to a further decline in the perceived relevance of the discipline. This would be retrogressive, as taxonomically intelligent information services depend on the expertise of the taxonomic community.

Acknowledgements

We appreciate the critical comments and influences provided to us by many colleagues along the way, and especially thank Donat Agosti, Adorian Ardelean, Jessie Kissinger, Phil Neal, Rich Pyle, Rod Page, Chuck Miller, Shauna Murray, Mindy Richlen, Sarah Roland, and Mitch Sogin.

References

Agosti, D., and N. F. Johnson. 2002. Taxonomists need better access to published data. Nature 417:222.

Biosis. 2005. Counts of new and changed animal names reported in Zoological Record, volumes 115–138. http://www.biosis.org/support/zr-changes/summary/

Bisby, F. A., M. A. Ruggiero, K. L. Wilson, M. Cachuela-Palacio, S. W. Kimani, Y. R. Roskov, A. Soulier-Perkins, and J. van Hertum, eds. 2005. Species 2000 & ITIS Catalogue of Life: 2005 Annual Checklist. CD-ROM. Species 2000, Reading, UK.

Büchen-Osmond, C. 2005. ICTVdB: The Universal Virus Database of the International Committee on Taxonomy of Viruses. http://www.ncbi.nlm.nih.gov/ICTVdb/INA

Chavan, V., N. Rane, A. Watve, and M. Ruiggiero. 2005. Resolving taxonomic discrepancies: Use of electronic catalogues of known organisms. Biodivers. Informa. 2:70–78.

Don, D. 1825. Prodromus Floræ Nepalensis; sive enumeratio vegetabilium, quæ in itinere per Nepauliam detexit atque legit F. Hamilton (olim Buchanan). Accedunt Plantæ a D. Wallich nuperius missæ. Secundum methodi naturalis normam disposuit atque descripsit D. D. London.

Dujardin, F. 1841. Histoire naturelle des Zoophytes—Infusoires. Paris.

Environment Australia. 1998. The Darwin Declaration. Australian Biological Resources Study, Environment Australia, Canberra.

Euzéby, J.-P. 2005. List of Bacterial Names with Standing in Nomenclature. http://www.bacterio.cict.fr/

Farr, E.R., and G. Zijlstra. 2005. Index nominum genericorum (plantarum). http://rathbun.si.edu/botany/ing/.

Froese, R., and E. Capuli. 2005. The Synonyms Table. http://www.fishbase.org/manual/fishbasethe_synonyms_table.htm

Geoffroy, M., and W. G. Berendsohn. 2003. The concept problem in taxonomy: Importance, components, approaches. *in* MOReTax: Handling factual information linked to taxonomic concepts in Biology (W. G. Berendsohn, ed.). Schriftenreihe für Vegetationskunde 39:5–14.

Godfray, C. H. J. 2002. Challenges for taxonomy. Nature 417:17–19.

Grimaldi, D., and M. S. Engel. 2005. Evolution of the insects. Cambridge University Press, Cambridge, UK.

Guiry, M.D., and E. Nic Dhonncha. 2005. AlgaeBase, version 3.0. Worldwide electronic publication, National University of Ireland, Galway. http://www.algaebase.org; searched on 28 May 2005.

Hebert, P. D. N., A. Cywinska, S. L. Ball, and J. R. deWaard. 2003. Biological identifications through DNA barcodes. Proc. R. Soc. Lond. B 270:313–321.

Herbert, K. G., N. H. Gehani, W. H. Piel, J. T. L. Wang, and C. H. Wu. 2004. BIO-AJAX: An extensible framework for biological data cleaning. Sigmod. Rec. 33:51–57.

Hobern, D. 2004. Should the TDWG names/concepts standard include BioStatus information? http://circa.gbif.net/Public/irc/gbif/dadi/newsgroups?n=dadi&a=re&art=25

Hopkins, G. W., and R. P. Freckleton. 2002. Declines in the numbers of amateur and professional taxonomists: Implications for conservation. Anim. Conserv. 5:245–249.

Knapp, S., E. N. Lughadha, and A. Paton. 2005. Taxonomic inflation, species concepts and global species lists. Trends Ecol. Evol. 20:7–8.

Koning, D., I. D. Sarkar, and T. Moritz. 2005. Taxongrab: Extractring taxonomic names from text. Biodivers. Informat. 2:79–82.

Lee, M. S. Y. 2000. A worrying systematic decline. Trends Ecol. Evol. 15:346–348.

Manaaki Whenua Landcare Research. 2005. http://www.landcareresearch.co.nz/

Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y.-J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C.-H. Ho, G. P. Irzyk, S. C. Jando, M. L. I. Alenquer, T. P. Jarvie, K. B. Jirage, J.-B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, and J. M. Rothberg. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380.

McNeill, J. 1997. Key issues to be addressed. *In* The new bionomenclature: The BioCode debate (D.L. Hawksworth, ed.). Biology International Special Issue 34:17–40.

Neave, S. A. 1939–1996. Nomenclator zoologicus; a list of the names of genera and subgenera in zoology from the tenth edition of Linnaeus, 1758, to the end of 1935 (with supplements). Zoological Society of London, London.

Page, R. D. M. 2005. A taxonomic search engine: Federating taxonomic databases using web services. BMC Bioinformat. 6:48.

Patterson, D. J. 2003. Progressing towards a biological names register. Nature 422:661.

Patterson, D. J., and J. Larsen. 1991. Nomenclatural problems with protists. Regnum Vegetabile 123:197–208.

Patterson D. J., D. Remsen, and C. Norton. 2003. Comment on Zoological record and registration of new names in Zoology. Bull. Zool. Nomenc. 60:297–299.

Polaszek, A., D. Agosti, M. Alonso-Zarazaga, G. Beccaloni, P. de Place Bjørn, P. Bouchet, D. J. Brothers, Earl of Cranbrook, N. Evenhuis, H. C. J. Godfray, N. F. Johnson, F.-T. Krell, D. Lipscomb, C. H. C. Lyal, G. M. Mace, S. Mawatari, S. E. Miller, A. Minelli, S. Morris, P. K. L. Ng, D. J. Patterson, R. L. Pyle, N. Robinson, L. Rogo, J. Taverne, F. C. Thompson, J. van Tol, Q. D. Wheeler, and E. O. Wilson. 2005. A universal register for animal names. Nature 437:477.

Pullan, M. R., M. F. Watson, J. B. Kennedy, C. Raguenaud, and R. Hyam. 2000. The Prometheus taxonomic model: A practical approach to representing multiple classifications. Taxon. 49:55–75.

Saarenmaa, H., 2002. Technological opportunities and challenges in building a global biological information infrastructure. (Pages 49–59) *in* Towards a global biological infrastructure (H. Saarenmaa and E. H. Nielsen, eds.). European Environmental Agency, Copenhagen.

Stein L. 2002. Creating a bioinformatics nation. Nature 417:119–120.

TDWG. 1985. Taxonomic databases working group. Minutes of the first meeting at the Conservatoire et Jardin Botanique, Geneva. http://www.tdwg.org/first_minutes.pdf

Thompson, C. 2005. The Diptera site. http://www.sel.barc.usda.gov/Diptera/diptera.htm

Thorne, J. 2003. Zoological record and registration of new names in zoology. Bull. Zool. Nomenc. 60:7–11.

Venter, J.C., K. Remington, J. F. Heidelberg, A. L. Halpern, D. Risch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Nealson, O. White, J. Petersen, J. Hoffman, R. Parsons, H. Baden-Tollson, C. Pfannkock, Y.-H. Rogers, and H. O. Smith. 2004. Environmental genome shotgun sequencing in the Sargasso Sea. Science 304:66–74.

Wheeler, Q. D. 2004. Taxonomic triage and the poverty of phylogeny. Phil. Trans. R. Soc. Lond. B 359:571–583.

Wilson, E. O. 2003. The encyclopedia of life. TREE 18:77-80.

First submitted 27 June 2005; reviews returned 17 August 2005; final acceptance 16 November 2005
Associate Editor: Rod Page