Zootaxa 1950: 153–163 (2008)
www.mapress.com/zootaxa/

ISSN 1175-5326 (print edition)
ZOOTAXA
ISSN 1175-5334 (online edition)

Principles for a names-based cyberinfrastructure to serve all of biology*

DAVID J. PATTERSON[1,3], SARAH FAULWETTER[2] & ALEXEY SHIPUNOV[1]

[1]*Biodiversity Informatics Group, Encyclopedia of Life, Marine Biological Laboratory, 7 MBL Street, Woods Hole, MA 02543, USA*
[2]*Institute of Marine Biology and Genetics, Hellenic Centre for Marine Research, 71003 Heraklion, Crete, Greece*
[3] *E-mail: dpatterson@eol.org*

** In: Minelli, A., Bonato, L. & Fusco, G. (eds) Updating the Linnaean Heritage:Names as Tools for Thinking about Animals and Plants. Zootaxa*, 1950, 1–163.

Table of contents

Abstract

The rapidly growing amount of biological data on the internet and the increasing need for large-scale analyses mandate improvements to the management of taxon-centric information. This information, traditionally managed by taxonomists, is now transforming into a web-based infrastructure. The complexity and narrative quality of the biological sciences require an information management framework that is sensitive to the scale, richness, character, and heterogeneity of the discipline. Given that the names of organisms offer us a nearly universal system for indexing biological data objects, a names-based cyberinfrastructure has the capacity to index the totality of available biological information and to aggregate taxon-centric data over a broad scale. In order to serve its role, this infrastructure should incorporate thirteen principles that are proposed here.

Key words: Biodiversity informatics, Taxonomic intelligence, Encyclopedia of Life, Data management, Taxonomy

Accepted by A. Minelli: 24 Oct. 2008; published: 5 Dec. 2008

Introduction

Shifts in the agenda of the biological sciences in the last two decades have been driven by diverse factors such as unifying molecular technologies, the challenges of climate change or, associated with the latter, the "*biodiversity crisis*" — the loss of biodiversity at all levels. In the latter area, both predictions and preparations for change will need analyses that integrate biospheric, economic, historical, social and geospheric information and on a scale that has previously not been considered. Biologists and conservationists need to synthesise different layers of understanding in order to "understand the whole" (Bisby 2000) and they require the availability of this information at an ever-increasing rate (De Carvalho *et al.* 2008). The extraction of understanding from the data will be empowered by an organisational framework that can interconnect biological information distributed in heterogeneous environments across the Internet. Ideally, as articulated by E.O. Wilson (2003), we should look to a future in which any piece of information on a species on the Internet will provide a gateway to all other information on the same species. With such interconnections in place, users will be able to index biological data objects, promote their atomisation into the smallest semantic parts and through those processes, create a vast communal pool of readily available taxon-centric data, making novel large-scale analyses possible.

Mayr (2004) correctly pointed out that biology is an unusual scientific discipline. The management of biological information will require solutions that are sensitive to the oddities of biology. Traditionally, biological knowledge has been catalogued and organised by taxonomists. From the time of Linnaeus — "*Filum ariadneum botanices est systema, sine quo chaos est res herbaria*" (Linnaeus 1751), taxonomy has effectively unifed biological knowledge and prevented its disarray. The declining numbers in the taxonomic community, as well as the growing deluge of biodata (termed "*the second bioinformatics crisis*" by Godfray, 2002), require us to come up with new ways of managing biodiversity information (Godfray 2002, Bortulus 2008). The vision presented here is one of a biological information management environment that embeds taxonomic and nomenclatorial thinking into the design of databases, data schemas, transfer protocols, applications, etc. with the intent of assembling an infrastructure capable of managing any piece of information of any type or size for any taxon.

Taxonomy as information management system relies on two elements that have long served the unique character of biology: names and hierarchies. Names are associated with most usable data objects and so are the common denominator of heterogeneous information coming from distributed sources. In the world of contemporary informatics, names can serve as metadata for all data objects that relate to taxa. That is, they can be used to form the foundation of an indexing system for all biology. Hierarchical arrangements, such as taxonomic, phylogenetic or any other kind of classification, specify relationships between elements by placing them into nested structures. They serve as ontologies which can be used to add higher levels of organization to the metadata and the data they include can be used to test hypotheses represented by the ontologies.

The use of names as metadata, of hierarchies as ontologies, and of both for data management are not usually considered to be part of taxonomy, and may run counter to principles of compliance with the nomenclatural codes. As an example, algorithmic indexing of digital data, such as the content of the Biodiversity Heritage Library (BHL, http://www.biodiversitylibrary.org) has to rely on name-recognition tools (e.g. Koning *et al.* 2005). Yet, many "names" in documents are obsolete or misspelled. If we are to index and recover all information, we need to catalog and cross-reference not only the code-compliant names but also the archaic and the misspellings. The implementation of this vision is not straightforward as a names-based information management has to overcome a number of problems, of which the most significant are the "many names for one taxon" (synonymy) and the "same name for many taxa" (homonymy) problems. Both confound the collation of all relevant information about the same species. In the case of synonyms, a simple search with one name will fail to find information linked to other names used for the same taxon, and in the case of homonyms, a search will draw together information on different, often unrelated, taxa. These problems have

always been at the mainstream of the descriptive taxonomy *sensu* Godfray (Godfray, 2002). A names-based cyberinfrastructure therefore needs to emulate the practices of taxonomists who have traditionally managed information in this area — that is, the infrastructure must be "taxonomically intelligent". Taxonomically intelligent names-based information management has an enormous potential for the biological sciences, especially if those developments are designed to allow machine-to-machine dialog through the use of globally unique identifiers, standardised data schemas, and interoperable data transfer protocols (Page, 2006).

The *Encyclopedia of Life* (EOL, http://www.eol.org) is the first major integrative project within biology that is explicitly based on these principles. To be able to fulfil its goal of delivering Web pages for every species it must be able to automatically aggregate taxon-centric information across the full spectrum of biodiversity. EOL relies on names-based information management. We have identified and discuss here thirteen features that we believe a taxonomically intelligent names-based cyberinfrastructure must have if it is to be effective for all types of organisms, and for all pieces of information, past, present and future.

The principles

Inclusive

A names-based infrastructure that is intended for managing information about any and all forms of life, must be designed to include all entities that satisfy any definition of "life" — whether viruses (and even prions), both types of prokaryotes, protists, plants, animals, or fungi. This requires that the infrastructure move in a direction opposite to the current trend of fragmentation into subdisciplines. This fragmentation led to similar but mostly independent codes of nomenclature, whereas the goal of biodiversity binformatics should be a single unified system. All codes seek to establish stability and to remove ambiguity in the use of names. They foster these goals within the jurisdiction of each code, but the independence of the codes can promote ambiguity when it comes to organisms that do not fit comfortably within one particular code (the ambiregnal names; Corliss 1995). The emerging infrastructure must not only apply to all organisms, but also to respect all code-based nomenclatural practices inclusive of the more innovative PhyloCode (Cantino & de Queiroz 2000). This code departs from traditional nomenclatural practices by seeking to regulate names that depict monophyletic and holophyetic clades by explicitly using phylogenetic principles. Irrespective of the logic by which they are derived, the names act as metadata and will be organized within an ontology. A names based infrastructure must accommodate such schemas if it is to serve the advocates of this phylogenetically motivated nomenclature.

Comprehensive

If a system is to be capable of indexing any biological data object, it must be capable not only of accommodating some information about organism, but all information on all organisms. That will require the architecture to include any identifier that has been used to assign a taxonomic context to a data object. In biology, most identifiers are scientific names, but the approach must also embrace vernacular names, and surrogates (such as culture isolate numbers, sample or specimen numbers) that are used in place of names. In the future, automated indexing tools will analyse electronic repositories, identify the labels, and use the label to link the data object with the taxon. The success of these tools will depend on how well they handle mis-spelled names, obsolete names, differently abbreviated names (and authority information), names that have been distorted through OCR (Optical Character Recognition) errors or Web algorithms (e.g. the Flickr machine tag format (http://www.flickr.com) removes spaces between names, creating "Iguanaiguana" from *Iguana iguana*). Auto-

mated name-recognition tools work through recognition of known names and/or the discovery of unknown names (Leary *et al.* 2007). Their functioning is facilitated through the assembly of a pool of all known names in all of their forms which can then serve as the basis for recognition algorithms, and for the improvements of name-discovery rules. The need for such a structure had led to the creation of the uBio NameBank (http://www.ubio.org), currently with over 10,000,000 names for 1,800,000 species.

Taxonomically intelligent reconciliation

The most widespread problem in the use of names for indexing purposes is that there are many different names and variants of names for the same species. The conventional taxonomic solution uses nomenclatural principles to select the correct name for the taxon, to which some or all of the other code-compliant names (synonyms) that have been applied to the same taxonomic concept can be linked. This logic lies behind the Catalog of Life Partnership compilation (CoLP, http://www.catalogueoflife.org). However, a nomenclaturally based solution such as this cannot form the basis of an indexing system because many names and variants of names which are associated with data objects are not code-compliant and so will be excluded. Moreover, the nomenclatural solution provides for the correct names at the present time, but cannot be retrospectively applied to many older documents. A purely nomenclatural approach can not serve well the needs of other major initiatives, such as the Biodiversity Heritage Library. An alternative solution, which is adopted here, is to catalogue all of the name strings that have been used for an entity and group them together within a reconciliation groups. The members of each reconciliation group contain all of the names that have been used for a given taxon. A query starting with any of the names in a group can exploit the reconciliation group to explode the query so that it uses all names. Reconciliation thus improves recovery of records, especially with older data (Table 1).

TABLE 1. Recovery of records from PubMed (http://www.ncbi.nlm.nih.gov/pubmed), JSTOR (http://www.jstor.org) and Google (http://www.google.com) with 5 of the 15 known related names for the red spotted newt from eastern North America.

Name	Year of first use	Items in PubMed	Items in JSTOR	Items in Google
Notophthalmus viridescens	1965	377	281	31,900
Diemictylus viridescens	1959	36	38	2,180
Triturus viridescens	1949	99	280	14,100
Diemyctilus viridescens	1965	1	3	105
Diemyctylus viridescens	1964	4	70	1,830

Any names that refer to different taxa (i.e. are homonyms) can be disambiguated through the use of reconciliation groups. Disambiguation requires copies of the same name string that refers to different taxa (such as *Aotus*, or *Peranema*) to be placed in different reconciliation groups and to be annotated with a flag that alerts users (whether people or machines) to a need for special attention.

Reconciliation groups contain names in several categories. Lexical variants are alternative spellings of the same term. An example might be different yet code-compliant spellings of the same name: such as *Gerardia paupercula var. borealis* (Pennell) Deam and *Gerardia paupercula* (A.Gray) Britton subsp. *borealis* (Pennell) Pennell, mis-spellings (*Eugelna* vs. *Euglena*), binomials with different endings (like *Pomatomus saltator* and *Pomatomus saltatrix*) and even abbreviations like *Camp rotu* (for *Campanula rotundifolia*) widely used in plant ecology. All are unarguably variations of the same name and so are objectively linked. Reconciliation

groups include so-called “objective synonyms” that include homotypic synonyms (*Pinus abies* Linnaeus and *Picea abies* (Linnaeus) H. Karsten), nomenclatural variants and combinations of names based on the same type material (e.g., *Pomatomus saltator*, *Temnodon saltator*, and *Gasterosteus saltatrix*), as well as subjective synonyms — names based on different types but accepted as synonyms in a particular treatment (e.g., *Dactylorhiza baltica* and *Dactylorhiza purpurella*). Because of their subjective component, reconciliation groups can be considered as a form of a taxonomic concept — a circumscription of the underlying biological meaning (see *Concept-capable* discussed below). Reconciliation groups must also extend to vernacular names because they too label data objects and in some environments are preferred over code-compliant names. Vernaculars will need to be placed within a linguistic, geo-referenced and script-based context given that the same names are often used for different organisms and the pattern of use depends on location. Reconciliation groups also include surrogates for names (such as culture identifiers, herbarium labels, nucleotide database IDs, etc.).

Names and their relationships with each other within the reconciliation groups can be annotated by flags to distinguish what kind of name the string refers to, the nature of the relationship among names, the provenance of assertions (Smith believes this is a junior subjective synonym of that), or to indicate the nomenclaturally correct name.

Taxonomically intelligent disambiguation

A names-based infrastructure must be capable of discriminating among the different uses of identical name strings for different taxa. Without this, automated systems have a high risk of confounding information on hemihomonyms — homonyms assigned to different taxa subject to different codes (Kluge 2000) like *Oenanthe* (the plant) with *Oenanthe (*the bird). One step for this class of homonyms is to mark all such names with code identifiers (as it is common practice in linguistics and philosophy). For homonyms falling under the jurisdiction of the same code (such as *Argus*, used for spiders, molluscs, birds and various insects) other means of disambiguation, such as providing the taxonomic context as discussed above, may be employed. These names should have a flag that alerts users and systems to the homonym problem. Any action involving one of these names must initiate a process of disambiguation that will lead to the association of a data object with the correct reconciliation group. Disambiguation is essential for automatic names-finding and indexing tools that will frequently encounter spelled-alike abbreviations (such as *C. marina*) that may refer to many different taxa. Name-recognition tools will require rule sets that can clarify the intention from the context in which the string appears. Scientific names can also be disambiguated by reference to broader taxonomic categories (*Peranema* Pteridophyta is not the same as *Peranema* Protista), by the naming authority (*Peranema* Dons vs. *Peranema* Dujardin), or by key words that associate with the target taxa (i.e. the co-occurrence of terms like “frond” or “spores”, or the names of other fern genera or species would indicate that the *Peranema* refers to the plant and not the protist). However, in many cases, the rules may be insufficient to disambiguate taxa, and in such cases, the indexing must be vetted and finalized through the community of experts to ensure that data objects are correctly classified. Vernacular names can be disambiguated by reference to scientific names that are used in conjunction with them, or through their linguistic or geographic context.

Concept-capable

Kennedy and co-workers (2005) argue that names are imprecise flags of “taxonomic concepts”. We may use the same name but have different opinions about what it refers to (e.g., does the name *Gorilla beringei* include the individuals that others refer to as *Gorilla beringei graueri* or not?). Taxonomic concepts refer to the scope

of application of the name of a taxon. This can be done through bibliographic references ("*sensu* Smith 1900"), references to specimens, or comparisons of different taxonomies that include or exclude other taxa. The challenges of concept management and the supremacy of concepts over names have led some to dismiss the value of a names-based infrastructure (Berendsohn 1995, Kennedy *et al.* 2005). Yet, a purely concept-based indexing system will be ineffectual because names are predominantly applied without clear indication of the concept to which they refer (Agnarsson & Kuntner 2007, Bortulus 2008). The best solution would be a marriage of a names-based with a concept-based management system. Taxonomic concepts can be incorporated within a names-based infrastructure in several ways. The inclusion of subjectivism within reconciliation groups offers one solution, multiple classifications (see below under "*Hierarchical structure*") offer a second, and placement of taxonomic concepts within a particular hierarchy offers a third.

Hierarchical structure

Classifications are important components of taxonomy because they represent hypotheses of the evolution of taxa or indicate relatedness. Hierarchical arrangements of taxa also provide a useful structure for biological data management (Kennedy 2003) as they may serve as ontologies defining relationships among (metadata) elements. Classifications can be exploited to disambiguate homonyms, browse content, drill towards more taxonomically precise groups or expand searches. They permit hierarchical aggregation of data, such that when a search is made on "Diptera", the settings can specify 'find me all data objects with the word "Diptera" associated with them', but also 'find me all data objects that carry the name of any fly or group of flies'. Hierarchical searches furthermore have significant value in being able to compare phylogenetic hypotheses by providing metrics of the consistency of data objects and their metadata with competing hypotheses.

Neither the entire species inventory nor the tree of life have been assembled, nor will they be. As a result, many different classifications will coexist and none of them is correct (Yoon & Rose 2001). A single, static classification will not be able to serve the needs of all users. Therefore, our management system has to be able to represent multiple, evolving hierarchies to reflect these different opinions about how organisms are related (see 'Phylogenetic'), eventually forming a graph of overlapping hierarchies (Kennedy 2003).

Phylogenetic structure

The hypothesis that all known life is inter-related through ancestor-descendant relationships remains unfalsified. That grounds the principle that seeks to incorporate our understanding of those relationships within our cyberinfrastructure. Closely related taxa share higher proportions of their genome, and can be expected to share a high proportion of their attributes. This provides a logical basis for the hierarchical organization within the system. It also allows us to infer and predict properties before they have been reported, an important feature as biology shifts towards datacentricity. The elimination and transformation of taxa that do not fully reflect our understanding of phylogenetic relationships characterizes the evolution of taxonomy. A cyberinfrastructure that can mimic the trend towards monophyly and holophyly will be more powerful and will gain acceptance from its user community. As noted under 'hierarchical', this trend should not be accompanied by a unitary point of view. Rather, the infrastructure can be designed to promote a process towards an architecture unified by phylogenetic relationships, and that process can be expressed in allowing stakeholders to change the composition and relationships of taxa.

Distributed organization

Information on names (whether lists of names, nomenclatural status, relationships) is located in many different on-line sources, all with their own specific purpose, taxonomic territory and user community (Table 2). There are many additional repositories of names and name associated information targeted on particular taxonomic groups (e.g., Antbase: http://www.antbase.org; CrustaceaNet: http://www.crustacea.net; Millipeet http://www.fieldmuseum.org/research_collections/zoology/zoo_sites/millipeet/), on particular habitats (such as the OBIS, ERMS, WoRMS and APHIA initiatives that address marine taxa), names relating to particular geographic regions, or lists holding vernacular names. This decentralisation of names providers has several advantages — it captures the enthusiasm of individuals by allowing them to identify closely with projects, it creates stability through redundancy, richness through diversity, relevance through purpose, and spreads the tasks among many players. A names-based infrastructure that can bring together the strengths of this distributed richness will outperform one that does not. On the other hand, information coming from different sources is usually heterogeneous in structure and value, so federation (integration of independent operations) is not straightforward. Creating a single point of access to this distributed information will help in organising information about, or attached to these names. This integration process can be addressed through devices like the Global Names Index (see *Up-to-date*, below) and by ensuring consistency through normalisation (see *Interoperability*, below).

Interoperability

Data flow between various names providers requires the adoption of standards, schemas and transfer protocols to facilitate the machine-to-machine dialog. Various standards exist in the field of biodiversity. The Biodiversity Information Standards group (TDWG, http://www.tdwg.org) is a key player in the development of standards for data exchange in different fields of biodiversity and promotes the deployment of Life Science Identifiers (LSIDs) to serve as globally unique identifiers (GUIDs) of taxonomic names. In order to capture the information which the data providers hold, a names-based infrastructure needs to be compliant with industry standards by adopting the current schemas, by serving LSIDs or other GUIDs through agreed data exchange standards. Furthermore, a names-based architecture should promote the usage of RDF (Resource Description Framework) formatted data and ontologies to facilitate semantic data exchange and retrieval. Many well-established databases would not be able to convert to new standards with ease, and in these cases, the solution will need to export and import data through “abstract layers” that transform data from one schema or format to another.

Up-to-date

Taxonomy is an evolving discipline. New taxa are continuously being discovered, new relationships are being described, and taxa are split, merged or renamed to reflect the most current knowledge about the evolution and relatedness of organisms. This continuously evolving knowledge has to be reflected by a dynamic names-based cyberinfrastructure. Excepting molecular biology, which involves the submission of published genetic sequences to central registries, the dissemination of biocentric information is not organised centrally. The Global Names Index, of which a prototype was established in 2008 is an emerging federated web services environment that dynamically interconnects an array of names partners. Names partners may include authoritative nomenclatural sources (such as ZooBank, the International Plant Index — IPNI, Index Fungorum, the Universal Virus Database — ICTVdB) or other repositories of authoritative information (such as the Catalogue of

Life Partnership). The names partners link to a common index through Web Services that automatically keeps the index appraised of changes in each participating database. Partners unable to provide appropriate web services can pass simple names lists into a hosting service that informs the index on their behalf. The common index provides a searchable and machine accessible environment that keeps all partners up to date on new names and associated metadata.

TABLE 2. A selection of on-line resources providing names information.

Project name	URL	Description
AlgaeBase	http://www.algaebase.org	Names of terrestrial, freshwater and marine algae
Deutsche Sammlung von Mikroorganismen und Zellkulturen (DSMZ)	http://www.dsmz.de/microorganisms/bacterial_nomenclature.php	Lists of names of eubacteria and archaebacteria that are compliant with the *International Code of Nomenclature of Prokaryotes*
Index Algarum	http://ucjeps.berkeley.edu/INA.html	Names of terrestrial, freshwater and marine algae
Index Fungorum	http://www.indexfungorum.org	Index of all code-compliant fungus names
Index Nominum Genericorum	http://www.botany.si.edu/ing	Compilation of generic names for organisms covered by the International Code of Botanical Nomenclature
Integrative Taxonomic Information System (ITIS)	http://www.itis.gov	Taxonomic information on plants, animals, fungi, and microbes, mostly of North America
International Plant Names Index (IPNI)	http://www.ipni.org	Names of genera and species of seed-bearing plants with their place of publication
List of Prokaryotic Names with Standing in Nomenclature	http://www.bacterio.cict.fr/allnames.html	Lists of names of eubacteria and archaebacteria that are compliant with the *International Code of Nomenclature of Prokaryotes*
micro*scope	http://microscope.mbl.edu	Information on the biodiversity of microbes
Nomenclator Zoologicus	http://uio.mbl.edu/NomenclatorZoologicus	Compilation of genera and subgenera in zoology from 1758 to 2004
The Catalogue of Life Partnership (CoLP)	http://www.catalogueoflife.org	An incomplete catalogue of all known species of organisms on Earth
uBio	http://www.ubio.org	Assembles all name strings ever used for organisms in literature and Internet, mainly for indexing purposes
Universal Virus Database (ICTVdb)	http://www.ictvdb.rothamsted.ac.uk	Approved virus names, linked to virus descriptions
World Register of Marine Species (WoRMS)	http://www.marinespecies.org	Comprehensive list of names of marine organisms, including information on synonyms
ZooBank	http://www.zoobank.org	Intended as the official registry of Zoological Nomenclature

Participation

The assembly and maintenance of a names-based infrastructure requires schemas, rules and algorithms to automate processes. Yet biology is not a "units and rules" science within which the totality can be derived from a sum of all of the parts, nor is it as 'rectangular' as informaticians might like it to be (where 'rectangular' refers to data in columns and rows). Biology has an inherent narrative component, and the elements of our understanding, whether the taxonomic perspectives, or the -ologies that transect the discipline, lack the atomic character of many other sciences, and so require an interpretative approach. The historical assembly of the science has often been by form of a social narrative in which personas have played a significant role in determining what is 'true'. An infrastructure for biology will not be one based on a few simple principles. Rather, it must handle data objects deriving from a complex, layered, inconsistent and sometimes unpredictable system. With tens of millions of names to manage, the nuancing of the system can only be achieved through active community involvement. Taxonomic experts who are willing to act as custodians for a clade will shoulder the responsibility of keeping the information up to date and of refining the crude algorithm-based approach to suit the nature of the discipline.

At another level, we can promote the evolution of the infrastructure so that it grows to become more appropriate to the task. By placing the concepts, algorithms, and tools into a communal open-source environment and by opening up content through APIs (Application Programming Interfaces) and other web services, we create the foundation for an evolutionary process to come into play and facilitate the emergence of cybertaxonomy as a cornerstone of the discipline. Furthermore, to encourage participation, an attribution system should be provided for all types of contribution to the system (e.g., provision of data, taxonomic editing of clades).

Authoritative

Despite the shapelessness of much of biology, there are better practices and poorer practices. Nomenclatorial aspects of taxonomic practices are usually regulated by the International Codes of Nomenclature following a framework of rules and recommendations that provide a certain level of structure and reliability to the subject. Nevertheless, the accumulation of data from heterogeneous sources will reveal errors and inconsistencies or even introduce new problems (e.g., algorithmically created errors, OCR errors, unicode conversion errors or erroneous "names" introduced by automated name-recognition tools).

Especially because of the peculiarities and widespread relevance of biology, and because of the dependence of a names-based architecture on taxonomy, it will be critical that devices are in place to facilitate continuing improvements in quality (accuracy, precision and completeness of the data environment). These will allow the system to evolve towards authority. Devices to support automatic quality control and quality assurance should be implemented, such as simple consistency checks, algorithms implementing nomenclatural rules, loops that return information from users to providers with the intent of improving fitness for purpose, clear indication of the quality status of a name (e.g., "vetted by expert", "coming from an authoritative source", "unverified status"), as well as devices that allow the hypotheses (such as the hypothesis 'Chromista') to be tested for consistency against bodies of all indexed data). Transparency and documentation of all elements will help the system to grow into a trustful source of information. No algorithm will ever be able to capture all deviations from general trends such that a system that combines algorithmic solutions to the challenges of scale, integration with clade custodianship by experts, together with devices to allow co-existence of multiple points of view, will more likely achieve the authority that we seek.

Scalable

One of the crucial requirements for the success of a names-based cyberinfrastructure is scalability. In the near future, the online availability of information and the increasing interconnection of data sources will result in rapidly growing numbers of names, relationships between them and data objects attached to them. As an example, new technologies of pyrosequencing have the capacity to generate millions of records of the diversity and abundance of species within ecosystems (Sogin *et al.* 2006)—within a matter of hours. The evolution of molecular technologies will result in growing accumulations of full genomes and community metagenomes. The Global Biodiversity Information Facility (GBIF, http://www.gbif.org) is currently setting itself the goal of indexing over a billion specimen records. Conservatively, we need to plan for 10^{12} data objects. The architecture of the underlying infrastructure must work with extremely large amounts of data, provide effective indexing and management of names, be stable, and permit effective discover and fast retrieval of date.

Conclusions

The EOL informatics group assisted by the PROPE-taxon initiative of the EU Network of Excellence MarBEF (Marine Biodiversity and Ecosystem Functioning, http://www.marbef.org) has been promoting the development of a names-based infrastructure with the properties described above. The components that are in place include repositories of names such as uBio's NameBank, sources of authoritative information — from the nomenclators to the aggregators, the adoption of standards (TDWG) and the emergence of a dynamic networking of names providers (GNI). The release of the Encyclopedia of Life in February 2008 exploited a prototype of this environment, proved that this approach is feasible and that it will contribute to a comprehensive and authoritative management of biological information at large. We are now in the process of building a system with the properties identified here.

Acknowledgements

The authors acknowledge support by the MARBEF EU funded Network of Excellence (contract no. GOCE-CT-2003-505446). Support was also received from ECOSUMMER (FP6-2004-Mobility-2-020501-2). The EOL team acknowledges support from the John D. and Catherine T. MacArthur and the Alfred P. Sloan foundations. Christos Arvanitidis and David Shorthouse are kindly acknowledged for valuable comments on this manuscript.

References

Agnarsson, I. & Kuntner, M. (2007) Taxonomy in a changing world: seeking solutions for a science in crisis. *Systematic Biology*, 56, 531–539.

Berendsohn, W. (1995) The concept of “potential taxa” in taxonomic databases. *Taxon*, 44, 207–212.

Bisby, F.A. (2000) The quiet revolution: biodiversity informatics and the internet. *Science*, 289, 2309–2312.

Bortulus, A. (2008) Error cascades in the biological sciences: the unwanted consequences of using bad taxonomy in ecology. *Ambio*, 37, 114–118.

Cantino, P.D., & de Queiroz, K. (2000) *PhyloCode: a phylogenetic code of biological nomenclature.* Available from: http://www.ohiou.edu/phylocode (accessed 20 June 2008).

Corliss, J.O. (1995) The ambiregnal protists and the codes of nomenclature: a brief review of the problem and of proposed solutions. *Bulletin of Zoological Nomenclature*, 52, 11–17.

De Carvalho, M.R., Bockmann, F.A., Amorim, D.S., Brandão, C.R.F., de Vivo, M., de Figueiredo, J.L., Britski, H.A., de Pinna, M.C.C., Menezes, N.A., Marques, F.P.L., Papavero, N., Cancello, E.M., Crisci, J.V., McEachran, J.D.,

Schelly, R.C., Lundberg, J.G., Gill, A.C., Britz, R., Wheeler, Q.D., Stiassny, M.L.J., Parenti, L.R., Page, L.M., Wheeler, W.C., Faivovich, J., Vari, R.P., Grande, L., Humphries, C.J., DeSalle, R., Ebach, M.C. & Nelson, G.J. (2008) Taxonomic impediment or impediment to taxonomy? A commentary on systematics and the cybertaxonomic-automation paradigm. *Evolutionary Biology*, 34, 140–143.

Godfray, H.C.J. (2002) Challenges for taxonomy. *Nature*, 417, 17–19.

Kennedy, J. (2003) Supporting taxonomic names in cell and molecular biology databases. *OMICS: A Journal of Integrative Biology*, 7, 13–16.

Kennedy, J.B., Kukla, R., Paterson, T. (2005) Scientific names are ambiguous as identifiers for biological taxa: their context and definition are required for accurate data integration. *In:* Ludäscher, B. & Raschid, L. (Eds.), *Data integration in the life sciences. Proceedings of the Second International Workshop 2005, San Diego, CA, USA.* Springer-Verlag, Berlin, pp. 80–95.

Kluge, N.J. (2000) *Modern systematics of insects. Part I. Principles of systematics of living organisms and general system of insects, with classification of primary wingless and paleopterous insects*. Lan', S. Petersburg, 332 pp.

Koning, D., Sarkar, I.N. & Moritz, T. (2005) Taxongrab: extracting taxonomic names from text. *Biodiversity Informatics*, 2, 79–82.

Leary, P.R., Remsen, D.P., Norton, C.N., Patterson, D.J. & Sarkar, I.N. (2007) uBio RSS: Tracking taxonomic literature using RSS. *Bioinformatics*, 23, 1434–1436.

Linnaeus, C. (1751) *Philosophia botanica in qua explicantur fundamenta botanica cum definitionibus partium, exemplis terminorum, observationibus rariorum*. G. Kiesewetter, Stockholm.

Mayr, E. (2004) *What makes biology unique?* Cambridge University Press, Cambridge, U.K., 246 pp.

Page, R.D.M. (2006) Taxonomic names, metadata and the semantic web. *Biodiversity Informatics*, 3, 1–15.

Sogin, M.L., Morrison, H.G., Huber, J.A., Welch, D.M., Huse, S.M., Neal, P.R., Arrieta, J.M. & Herndl, G.J. (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences*, 103, 12115–12120.

Wilson, E.O. (2003) The encyclopedia of life. *Trends in Ecology & Evolution*, 18, 77–80.

Yoon, N. & Rose, J. (2001) An information model for the representation of multiple biological classifications. *In:* Alexandrov, V.N., Dongarra, J.J., Juliano, B.A., Renner, R.S. & Tan, C.J.K. (Eds.), *Computational Science - ICCS 2001: International Conference, San Francisco, CA, USA.* Springer-Verlag, Berlin, pp. 937–946.