Zootaxa 1950: 5–8 (2008)
www.mapress.com/zootaxa/

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

Actual usage of biological nomenclature and its implications for data integrators; a national, regional and global perspective*

CHARLES HUSSEY[1], YDE DE JONG[2] & DAVID REMSEN[3]

[1]*Natural History Museum, Cromwell Road, London SW7 5BD UK. E-mail: c.hussey@nhm.ac.uk*
[2]*University of Amsterdam, P.O. Box 94766, NL-1090 GT, Amsterdam, The Netherlands. E-mail: yjong@uva.nl*
[3]*Global Biodiversity Information Facility, Universitetsparken 15, DK-2100 Copenhagen, Denmark. E-mail: dremsen@gbif.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.

Abstract

Biological names play an important role in resource identification and as anchors for all sorts of associated information. This is borne out in ever-expanding online resources but the ways in which names are stored and presented give rise to challenges and pitfalls that can lead to missed or misinterpreted information. These resources must serve a variety of users and keep abreast of changes in nomenclature and systematics. Observations on the use of biological names are presented and some solutions to the challenges are offered.

Key words: Biodiversity, Digital data bases, Nomenclators

Accurate identification of organisms and correct use of biological names is essential in order to apply correct measures in the fields of conservation and to control pest and disease causing organisms. As has been pointed out by Grimaldi and Engel (2005) "*All accumulated information of a species is tied to a scientific name, a name that serves as a link between what has been learned in the past and what we today add to the body of knowledge*". While the veracity of the statement holds true, the nature of taxonomy and nomenclature present significant obstacles to taking advantage of this universal link between a taxon name and the accumulated information.

Efforts to mobilise biodiversity information have now yielded significant online resources, and these are set to grow enormously in the future. The Internet is revolutionising accessibility but also creates its own set of obstacles to discovery and retrieval of information based on taxonomic names. At the time of writing, the Global Biodiversity Information Facility network (http://www.gbif.org) has mobilized nearly 150 million collection and observation records from nearly 3,000 individual datasets. The Biodiversity Heritage Library (http://www.biodiversitylibrary.org) has recently passed the 7 million page mark toward its goal of digitizing an estimated 2–3 million publications relating to species. The National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov) stores tens of millions of gene sequence relating to more than three hundred thousand taxa. The Biodiversity Information Standards (http://www.tdwg.org) website lists 592 different biodiversity informatics projects, all of which are mobilizing, serving, integrating and exchanging species information. Each of these resources shares a common dependence on taxon names to provide the species context to the associated information.

There are also numerous information resources relating to the compilation and reconciliation of taxon names. The authors have each been engaged in long-term projects that involve collating lists of taxonomic

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

names from disparate sources and making these accessible through the Internet (National Biodiversity Network Species Dictionary http://www.nhm.ac.uk/nbn, Nature Navigator http://www.nhm.ac.uk/naturenavigator, Fauna Europaea http://www.faunaeur.org, GBIF Electronic Catalogue of Names http://www.gbif.org/prog/ecat, uBio http://www.ubio.org).

In our experience, taxon names present a range of challenges that must be addressed in order to realize their potential as useful data discovery devices. Among the most fundamental are conceptual separations that tend to be blurred and can hinder communications. In particularly there is little appreciation of the difference between taxonomy and nomenclature. This impacts upon concepts as simple as what is meant by the term "correctness" as it relates to taxon names or as fundamental as the term "name" itself. All taxa are referred to by a name but not all names refer to currently recognized taxa. A "correct name" in the taxonomic sense may refer to the valid or accepted name for a taxon, whilst in the nomenclatural sense it may refer to the fitness of the name relative to the codes of nomenclature. The scope of a name can change over time, as the result of improved knowledge, or through differing opinions amongst experts. Strictly speaking, names should be replaced by taxon concepts, which are names linked to explicit usage. This requires a record of the source used in making a taxonomic determination – metadata that is most often missing from a species observation record.

In these days of internet resources, this lack of distinction can lead to confusion, relevant information being missed, or information being incorrectly associated and returned in a search result. Among the challenges are the following:

1. Matching a name entered via a search to a name actually recorded within a data repository. Computers excel at comparing text strings but differences in how names are recorded can result in both false negative and false positive returns.
2. There is, as yet, no comprehensive catalogue or index representing all the taxon names that exist, nor any accurate measure as to the true number of names.
3. It is becoming apparent that copies of some datasets are included in other resources, with or without the agreement of the original data owners. Any errors in the original dataset may therefore be carried through to the derivative resource and may persist even after the original has been corrected.

Bearing in mind that there are around 1.75 million described species (Wilson 2003), it is noteworthy that the number of names assembled by uBio within their *NameBank* currently amount to 11 million distinct name records. This is due primarily to the fact that *NameBank* records distinct verbatim name strings (name+authorship) combinations and reconciles these to a single logical group. The Catalogue of Life (http://www.catalogueoflife.org), on the other hand, starts with quality lists that have been reconciled to single representative name records and therefore, the 2008 edition, representing over 1.1 million species, has a total of 2 million name records. In the United Kingdom, which has around 80,000 species (excluding bacteria and viruses), the NBN Species Dictionary holds 245,000 name strings.

These counts include representational forms of both taxonomically and nomenclaturally valid and invalid names. Many of these will be obsolete names, subsequently made synonyms of current names. In addition, a single name may present a wide range of variability in how it is actually recorded within a dataset. The name may be non code-compliant, have a wrong endings to the species epithet, or be simply misspelled. Variation may also occur in how authorship is represented (such as abbreviations and inconsistent use of diacritical marks). Such variation presents challenges in federated data environments where inconsistency is the rule.

Thus a single taxon name in the more traditional nomenclatural sense may be represented by many small lexical variations of that name, as they have been discovered within biodiversity resources. Nonetheless, the 11 million records within the uBio *Namebank* represent over 4 million distinct taxon names if authorship is no longer a factor. Interestingly there are over 4.8 million distinct name strings, based on this same definition, within the current GBIF indices, which can be reconciled to 3.4 million distinct names. Of these, only 11% are listed in the *Catalogue of Life* and only about 25% are believed to overlap those found within the *NameBank*.

This serves to show the scale of one of the challenges facing biodiversity informatics.

Recording names in use, even where they are erroneous, enables query expansion: provided that effort is put into mapping these names to their accepted forms. Mapping of names can, to a certain extent, be automated: uBio and GBIF, for instance, have developed their *LexMapper* algorithm to handle this. Older names will increasingly need to be tracked, now that specimen collections are being digitised, as well as the historic literature. Homonyms represent an unquantified but significant issue, even within a single biological kingdom, and become even more of an issue when resources span several kingdoms. In the process of digitising *Nomenclator Zoologicus*, 21,000 homonym groups were identified (Remsen et al. 2006). GBIF is developing an All Genus Index (AGI) that should identify all genus-level homonyms (Remsen & Patterson 2007).

Storing, comparing, exchanging and searching for taxonomic names and classification schemes also present challenges. Search portals usually offer the facility to search using the genus or species epithet. However, name strings can contain up to 14 words in the case of plant hybrids. Because database searches rely on string matching, variants in spelling (such as presence or absence of diacritical marks) can lead to missed records – unless such variants have been mapped to accepted forms.

Various data models and exchange standards have been developed over the years (ABCD, Berlin Taxonomic Information Model, Darwin Core, EDIT Common Data Model, Nomencurator, Taxonomic Concept Transfer Schema) to cope with biological names and classifications. It is possible to use these schemas to wrap name data to common formats, even if the underlying database has a unique structure. It should be borne in mind, however, that many of the data providers, particularly those involved in local and national recording schemes, may not be willing or able to use complex systems and, instead, often record and present data using simple spreadsheets or documents.

Biologists look to nomenclators and taxonomic indexing services for help in checking current names and their authorities, which is only possible if synonymies are included. But users will also include conservationists, developers and planners, local and national government, environmental agencies, biological recorders and members of public, who may have different needs. For instance, biological recorders require the inclusion of recording aggregates (an amalgam of species that are difficult to identify in the field) and wish to record against names that they are familiar with. Many users, who are not practising taxonomists, are not concerned with the niceties of nomenclatural and taxonomic rules, such as the use of subgenera and authorities – they just want a reliable name! Some sectors (e.g. birds, butterflies, mammals) routinely use common names. Informal names are also helpful for higher taxonomic groupings. Even biologists will be unfamiliar with names of genera, families and orders outside of their own speciality and it can greatly help if search results assign each scientific name to a familiar higher grouping.

There are numerous initiatives at national level, fewer at regional level and even fewer resources at global level. National coverage, both in terms of expertise and content, is uneven. There are numerous instances where data exists but are yet to be made accessible. Whilst a single checklist can achieve consistency, through being based upon a single taxonomic opinion, when datasets are assembled from multiple sources, these sources may employ different classifications and synonymies. Often, however, there is a preferred classification for a taxonomic group at a national level. The correspondence between vernacular names and scientific names may differ between countries and even the accepted scientific name for a species can vary. Equivalencies can be determined by assigning Globally Unique Identifiers (GUIDs) to taxa. It is, however, important that systems are able to allow for and support different taxonomic opinions. There also needs to be an effective exchange of information between national, regional and global initiatives. In that way, new occurrence records can be fed upwards and changes to nomenclature can be fed back to biological recorders.

What is needed is sustainable, long-term, initiatives that will deliver maintained taxonomic indexes and nomenclators. Whilst it is possible that the Lifewatch (for Europe) and Encylopedia of Life projects (http://www.lifewatch.eu/, http://www.eol.org), together with GBIF, will provide high-level access to data, the challenge is to secure support for the hundreds of individual data contributors. All resources, whether nomenclators or species inventories, should be kept abreast of changes: in order to be able to gauge whether a name is current and also whether a species occurrence is current. This requires continuous effort, and it is not easy to secure funding for this sort of activity. To take things forward, more attention should be given to the mapping of obsolete and malformed names to code-compliant accepted names, to flag the status of names, and to capture vernacular names. Development of a management classification will help ensure that consistent results are returned from searches across distributed datasets. Authorities are necessary to give attribution to a name, but the abbreviated form in which they are presented (in both botany and zoology) does not enable the determination of the underlying bibliographic reference. The increasing availability online of scientific literature should be complemented by a resource that not only links species names to their original description, but does the same for species recombinations (comb. nov.). It is to be hoped that use of GUIDs will become commonplace; with a management system that resolves multiple GUIDs that may get assigned to a single taxon concept. It is important that the provenance of datasets is indicated whenever records are displayed or downloaded. Attribution also provides welcome acknowledgement of the work of data providers, many of whom work on a voluntary basis. Above all, it is at the human level that action is required. Action to promote best practice in the use of names. Also action to mobilise the biological community to assist with error detection and correction, and to both share and consolidate resources, in order that the current duplication of effort may be reduced.

References

Grimaldi, D. & Engel, M.S. (2005) *Evolution of the Insects*. Cambridge University Press, Cambridge, 786 pp.

Remsen, D.P., Norton, C. & Patterson, D.J. (2006) Taxonomic informatics tools for the electronic *Nomenclator Zoologicus*. *Biological Bulletin*, 210, 18–24.

Remsen, D.P. & Patterson, D.J. (2007) *Building an index of all genera: A test case in interchange*. Available from: http://www.tdwg.org/proceedings/article/view/265 (accessed 6 September 2008).

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

Additional sources

Anonymous (2005) *Taxonomic Concept Transfer Schema*. Available from: http://tdwg.napier.ac.uk/ (accessed 6 September 2008).

Berendsohn, W.G. (page editor) (2007a) *Access to Biological Collection Data (ABCD) Schema*. Available from: http://www.bgbm.org/tdwg/codata/schema/ (accessed 6 September 2008).

Berendsohn, W.G. (2007b) *The Berlin Taxonomic Information Model*. Available from: http://www.bgbm.org/biodivinf/Docs/BGBM-Model/default.htm (accessed 6 September 2008).

Biodiversity Information Standards (2007) *Darwin Core*. Available from: http://www.tdwg.org/activities/darwincore/ (accessed 6 September 2008).

EDIT [European Distributed Institute of Taxonomy] (2008) *Common data model*. Available from: http://dev.e-taxonomy.eu/trac/wiki/CommonDataModel (accessed 6 September 2008).

Ytow, N., Morse, D.R. & Roberts D.McL. (2001) Nomencurator: a nomenclatural history model to handle multiple taxonomic views. *Biological Journal of the Linnean Society*, 73(1), 81–98. See also http://www.nomencurator.org/ (accessed 6 September 2008).