

## ZooBank: reviewing the first year and preparing for the next 250

RICHARD L. PYLE<sup>1</sup> & ELLINOR MICHEL<sup>2</sup>

<sup>1</sup>*Department of Natural Sciences, Bishop Museum, 1525 Bernice St., Honolulu, Hawaii 96817, USA*

<sup>2</sup>*International Commission on Zoological Nomenclature, c/o The Natural History Museum, Cromwell Road, SW7 5BD London, U.K.*

### Abstract

The ‘Linnaean Enterprise’, begun by Carolus Linnaeus in the mid-eighteenth century, stands as one of the most enduring (and arguably among the most important) scientific endeavors in human history: the quest to catalog all living species. During the 250 years since, this Enterprise has expanded in zoological content through a series of initiatives, including the Zoological Record, Sherborn’s *Index Animalium*, von Schulze’s *Nomenclator Animalium* and Neave’s *Nomenclator Zoologicus*. More recently, with the advent of computers and the internet, ambitious initiatives such as the Catalog of Life and Encyclopedia of Life have made great strides towards realizing Linnaeus’ original vision. In 2005, the ICZN Secretariat and Commissioners took one more step towards achieving this grand endeavor by proposing “ZooBank” as a web-based registry of zoological names and nomenclatural acts.

The ZooBank web site was launched as a prototype on January 1st, 2008, coinciding with the 250th anniversary of the official start of Zoological Nomenclature. At its launch, the ZooBank registry included all 4,819 names established in the 10th edition of Linnaeus’ *Systema Naturae*, as well as five new fish species names established in an article published concurrently with the launch of ZooBank. ZooBank is not intended to replace existing nomenclatural catalog databases, and it makes no assessment or judgment of the taxonomic content of any published work. ZooBank assigns unique registration identifiers in four information domains of relevance to the ICZN *Code*: nomenclatural acts (including new names and other acts that affect existing names), publications, authors, and type specimens. These identifiers are envisioned as pointers to authoritative information concerning zoological nomenclature and are expected to become integral to current and future efforts to index taxonomic content. The complete implementation details of the ZooBank registry are currently being discussed, developed, and tested, with involvement from ICZN Commissioners, nomenclatural data managers and the taxonomic community at large. Many questions concerning technical implementation details, content sourcing and prioritization, information quality standards, and scenarios for mandatory registration are open to discussion. As much as there is an urgent need to answer these questions soon, there is also the need to “get it right”, ensuring a solid foundation for the next 250 years of zoological taxonomy.

### Introduction

*Completing the Linnaean Enterprise*

With the publication of *Systema Naturae* in 1758, Carolus Linnaeus launched a system for naming and classifying animals that would endure for the next two and a half centuries, right up until today. Throughout this amazingly long (in the context of scientific methodology) span of time, the ‘Linnaean Enterprise’ of establishing ‘sense and stability’ to animal diversity has expanded far more in content, than it has in its basic goals, methods and vision. This is not in any way a failure by the taxonomic community to evolve over the years, so much as it stands as testament to the deep understanding and appreciation Linnaeus had for the natural living world. But Linnaeus was fortunate, in some sense, to have lived at a time when the entirety of zoological diversity could be represented in a single paper-printed volume. The names he gave the species he recognized were unique and unambiguous. He could not have known that, 250 years hence, despite a continued and accelerating process of species discovery, humanity would still not know the true extent of animal diversity even within an order of magnitude. He would have been irritated to note that his system of ordering names, which worked so well for a few thousand names in a single work, began to fray at the edges when information became dispersed, when zoologists continued to add to the system in different languages, from distant corners of the earth, with access to different subsets of the existing system of names. As the literature on species descriptions grew, the challenge of keeping it accessible and regularised was met by a few visionary biologists.

The Linnaean Enterprise was greatly expanded in the first half of the nineteenth century when Albrecht Günther founded the *Zoological Record* in 1834 as an effort to catalog all publications in zoology. Since its inception, the *Zoological Record* has not only archived the past accomplishments of zoological research, it has provided a model for biodiversity informatics, in that it integrates disparate fields of biology and bridges them through the “common denominator” of all biological sciences: scientific names. The sweeping ambition of *Zoological Record* has been a remarkable success story, providing a critical tool that has improved biologists’ ability to synthesize information of all kinds. It continues today, engaging a team of biologists to catalogue publications in over 5000 serials and other publications, entering 75,000 records each year. Over 3.8 million records can be found through *Zoological Record*. Although this gives access to the majority of zoological publications, it is estimated that there are still a significant proportion that escape this catalogue as they are published in obscure sources.

By the turn of the 19<sup>th</sup> century it was clear that there was a need to bring order to this deluge of biological information by providing a the backbone of a complete catalogue of scientific names, a nomenclator, to act as an authoritative source for the published origin of the names. Early attempts at nomenclators had been made by Agassiz (1848), Marschall (1873) and Scudder (1882), among others; however, they were insufficient and not maintained. In 1902 Charles Davies Sherborn published the first installment of what has become the most ambitious listing of all animal names to date, both fossil and recent, entitled *Index Animalium*. He was the right man for this job, as he was an exacting cataloguer, based at an institution with a superlative library of natural history (the then-named British Museum (Natural History), London) where he could check each reference personally. He had three unambiguous aims: (a) to provide a complete list of all generic and specific animal names, (b) to provide a reference for each description, and (c) to give an exact date for each page quotation (Sherborn, 1902 p. vi). His first volume of 1200 pages treated 60,000 species which he had directly checked in 1300 references that dated from 1758-1800; this took him twelve years, including a period of ill health (which he

carefully documents as totalling three years' break from the intensive work). He worked on this catalogue for 31 more years, producing eleven volumes, commenting that 'for accurate work it is necessary for the student to verify every reference he may find; it is not enough to copy from a previous author; he must verify each reference itself from the original. Bad work, for which there is little excuse, is only too common.' (Sherborn, 1932). *Index Animalium* is still seen as an essential resource today, with its on-line version receiving approximately 1150 unique visitor hits per month. This monumental work has justifiably earned Sherborn the title of "The Father of Biodiversity Informatics".

Sherborn's masterwork was arranged by species. This, for example, is the listing for our own taxon:

**sapiens** Homo, Linnaeus, Syst. Nat., ed. 10, 1758, 20 ; ed. 12, 1766, 28 ; *varr.* ferus, americanus, europaeus, asiaticus, afer, monstrosus.

His reasons for this arrangement have at their core the understanding that his work was a resource for the bibliographic origins of scientific names, which should remain independent of decisions on taxonomy. A nomenclator provides access to the data needed to determine priority and avoid homonymy, or duplication, of the same name for different taxa. Sherborn's clearly stated thoughts on this (1921, p. ix) were that:

1. No synonymy of species is attempted: that depends on the idiosyncrasy of the systematist.
2. Any attempt at specific synonymy would be opposed to progress, as experience shows that vast changes may take place in a single year.
3. An arrangement under species permits of a generic synonymy, for by running the eye down the second column of the printed work, it will be possible to ascertain the various generic names with which a particular species name has been connected."

However, a reliable nomenclator was also needed for generic (and sub-generic) names; homonymy is an equally insidious problem for these names, and the generic level is often the point of entry for taxonomic questions. Two independent efforts to redress this were published in the early part of the twentieth century. The first of these, *Nomenclator Animalium generum et subgenerum*, was published in 1926 by a widely respected invertebrate zoologist Franz Eilhard von Schulze and included generic animal names from Linnaeus. Work on *Nomenclator Animalium* continued through 1954 by W. Kükenthal and K. Heider. Another approach to the same goal was made by Sheffield Airey Neave, with his *Nomenclator Zoologicus*, which now catalogues the bibliographic origin of the 340,000 genera and sub-genera described from Linnaeus 1758 to 1994 (Remsen, et al. 2006). This work is available on-line and heavily used, drawing approximately 2300 searches per month from 19,000 unique users per year. Neave (1939, p. v) explained in the Forward to what was to become an 11 volume work, that in 1934 "systematists found themselves in great difficulties for lack of up-to-date information relating to generic names, largely owing to the fact that no index to new generic names had been published since 1910, and that the *Nomenclator Animalium generum et subgenerum* published by the Prussian Academy of Science was neither complete nor up-to-date, besides being very costly." Thus, this second effort at a generic nomenclator was undertaken for scientific, economic and, we might guess, a sliver of nationalistic reasons.

With the advent of computers and the internet came much more powerful ways to organize and access information. The Linnaean Enterprise expanded into the electronic domain early on in the history of the PC revolution, with scientists and amateurs immediately building taxonomic databases of varying degrees of scope, completeness and reliability. However, these efforts remained disconnected and inaccessible. The advent of internet connectivity and expanding vision sowed the seeds for ambitious Linnaean projects of an altogether grander scale such as the Catalogue of Life ([www.catalogueoflife.org](http://www.catalogueoflife.org)), and the Encyclopedia of Life ([www.eol.org](http://www.eol.org)). These initiatives aim to bring together taxonomic, and other relevant biological information, on recognized (or valid) taxa of living metazoans with the ambitious aim to eventually construct largely complete information sources. However the accompanying nomenclatural information is included largely as a by-product of the taxonomy, without a mandate to ensure that names are ICZN *Code*-compliant, and their authority does not include a legislative body on nomenclature. As so succinctly expressed by Sherborn, above, taxonomy needs a robust source for nomenclatural information that is independent of taxonomic judgments. Names come in and out of valid use depending on the perspective of the taxonomist making judgments on synonymy, however, only names that have been published according to the rules of the ICZN *Code* are available.

### *Linking Biodiversity Information Through Names*

Names are the natural organizing linkage for biological information, as language is the natural structure for thought and communication. Robust names with unambiguous meanings are important not only for taxonomists to recognize which units they are describing, but also for all other users of biological information, including people working in agriculture, human and veterinary medicine, conservation, ecology, molecular biology, law and policy. In fact, any human concern that involves the living world should ideally have an unambiguous nomenclature, however work involving science and its products may have a more explicit justification for unambiguous names than daily concerns. Linnaeus recognized this as a particular need for communication in science, as it is more international, more nuanced in detail and with greater consequences for misinterpretation than many other kinds of communication. This stimulated him to develop his binominal system. In the first edition of the ICZN *Code* (1961, p. iv), J. Chester Bradley said “Ordinary languages grow spontaneously in innumerable directions; but biological nomenclature has to be an exact tool that will convey a precise meaning for persons in all generations”.

Having the tool of an (ideally) exact, unambiguous nomenclature with 250 years of scientific legacy linked to the names gives us powerful leverage on information. The power of information technologies allows us to expand the vision of *Zoological Record*, which worked well even as print on paper, to make an order of magnitude more information available to all users, and make it easier to organize once it is in hand. This can only improve scientific practice, which builds incrementally on past knowledge.

### *The Role of ICZN and ZooBank*

If names are the logical links for scientific information about animals, we need to ensure that

those links are as robust and unambiguous as possible. As early as the mid-nineteenth century it was realized that without guiding rules and the cooperation of the taxonomic community, zoological nomenclature would falter, with consequent serious loss of access to information. A group of concerned luminary zoologists headed by Hugh Strickland and including Charles Darwin (originator of the theory of evolution), Richard Owen (renowned anatomist and founder of the Natural History Museum, London) and John Westwood (first Chair in Entomology at Oxford), set out a code of rules for the scientific naming of animals. Their body was the precursor for what is now the International Commission on Zoological Nomenclature (ICZN), an international organization of distinguished zoologists whose mandate is to continue to formulate and update rules for naming animals, and to adjudicate in situations of confusion or dispute. These rules are published in the ICZN *Code of Nomenclature* ('the *Code*', currently in its 4th Edition, 1999, and online [www.iczn.org](http://www.iczn.org)) with supplemental amendments and declarations. By ensuring that nomenclature is applied in a globally consistent way, the ICZN provides continuity both for new species discoveries and for the correction of errors and inconsistencies in past works. The current ICZN mandate also includes developing tools for making nomenclature accessible through modern technology.

Having a complete listing of existing available names, exposing and preventing further homonymy were central aim of the authors of the great nomenclators of the past; used appropriately, bioinformatics will make this task much easier to accomplish. In 2005, the ICZN Secretariat and Commissioners initiated the process of achieving this goal by proposing "ZooBank" as a web-based registry of zoological names and nomenclatural acts (Polaszek, et al., 2005). A series of meetings among the Commissioners and presentations to the taxonomic community allowed input in the initial model for ZooBank. ZooBank is not intended to replace existing nomenclatural catalog databases, and it makes no assessment or judgment of the taxonomic content of any published work.

### **Implementing ZooBank: "The Devil Hides in the Details"**

There are two aspects to implementing ZooBank: technical, and policy. For the most part, the technical implementation is relatively straightforward. Policy decisions, however, are much more subtle, and have proven to require more careful consideration. ZooBank derives its legitimacy through the ICZN, the widely recognized international body charged with maintaining the rules of zoological nomenclature. As such, policy decisions related to ZooBank are at the collective discretion of the ICZN Commissioners. In August 2008, in connection with the Symposium for which this article is written, the ICZN held a meeting in Paris that was attended by 13 Commissioners, with eight external professional participants. A full day of the three-day meeting was devoted to presentations and issues related to ZooBank, and during that meeting a ZooBank Committee was formed to move forward with discussions and recommendations for ZooBank development and design (Pyle, 2008). In the days that followed, a series of related meetings addressed the role of ZooBank in the broader context of emerging global taxonomic data infrastructure, as well as for potential data content providers. One of the clear themes that emerged from those meetings was the sense that, while the grand vision of ZooBank is easy to articulate, the specific implementation details are what require the most careful thought and consideration. The remainder of this article provides an overview and brief discussion of some of those considerations.

### *Initial Launch and Early Development*

The first public implementation of ZooBank was launched at midnight, GMT, on January 1<sup>st</sup> 2008 – exactly 250 years to the day after the official start of zoological nomenclature (i.e., the date officially fixed for the publication of *Systema Naturae*). At that time, the content of ZooBank (zoobank.org) included 4,819 names established in *Systema Naturae*, as well as five new species of damselfish published concurrently with the launch of ZooBank (Pyle *et al.*, 2008).

From its launch, through to the time of this writing, the public ZooBank web site only allows searching of and read-only access to existing ZooBank entries. Active development on a web-based registration interface over the past year has resulted in nearly a thousand additional names from over 3,500 publications, and over 4,000 authors. Two grants from GBIF (Global Biodiversity Information Facility) and one from TDWG (the biodiversity informatics standards body) have supported the implementation and content expansion process, and at least two journals (*Zootaxa* and *ZooKeys*) are proactively including ZooBank registration as part of their ongoing publication process.

Development and testing of web-based interfaces to allow addition of content and editing of existing content continues, in preparation for a wider launch to allow more active and direct participation by a broader set of practicing taxonomists.

### *Globally Unique Identifiers*

One of the core aspects of registration is the assignment of a persistent Globally Unique Identifiers (GUIDs). In a sense, the names themselves have served the function of identifiers from the perspective of humans communicating with humans; but due to changing combinations, alternate spellings both mandated by the *Code* (gender agreement and emendations) and unintentional (misspellings and typographical errors), as well as homonymy (both within zoology, and among names governed by other *Codes* of Nomenclature); these names are neither persistent, nor unique enough to serve the function of a GUID in the context of electronic information management.

ZooBank is following the lead of GBIF and TDWG in implementing Life Science Identifiers as the GUID for ZooBank registration entries. LSIDs conform to a standard format consisting of minimally five components, each separated by a colon (:). The first two components, “urn:lsid”, are common to all LSIDs, and simply establish the LSID as a “uniform resource name” (urn), and an LSID. The third component represents the “authority identifier”, which usually takes the form of a web domain name. As such, all ZooBank LSIDs incorporate the text, “zoobank.org” as the authority identifier. The fourth component of an LSID is the “namespace identifier”. In the case of ZooBank, this corresponds with the “domains” of data that will be registered in ZooBank. These are discussed in more detail in the following section, but the four namespace identifiers already implemented in ZooBank are “act” (for nomenclatural acts), “pub” (for publications), “author” (for zoological authors), and “specimen” (for primary, or name-bearing, type specimens).

The fifth component of an LSID is the “object identifier”, which is simply a number or other identifier that is unique within a particular namespace of a particular authority. According to the LSID specification, the combination of authority identifier, namespace identifier, and object identifier must collectively be unique on a global scale. However, the object identifier component of ZooBank LSIDs is itself a globally unique “Universally Unique Identifier” (UUID). UUIDs are 16-byte (128-bit) numbers, usually represented by hexadecimal characters (the numbers 0-9 and the letters A-F) in the format of 8 characters, a hyphen, three sets of four characters, each separated by a hyphen, followed by another hyphen and an additional 12 characters. Some unappealing aspects of using UUIDs in this context are that they are unpleasant to look at, nearly impossible for most mortals to memorize, and take a lot of space when typed out in their entirety. However, GUIDs in general, including ZooBank LSIDs in particular, are not intended to be optimized for human consumption and interpretation. Indeed, the taxon names themselves serve this function rather effectively (as they have for the past 250 years) – given the ability of human brains to resolve any ambiguity from homonymy or alternate spellings when provided sufficient context. Rather, ZooBank LSIDs (like all GUIDs) are optimized for consumption and interpretation by computers. Moreover, by using the globally-unique UUIDs as the object identifiers within LSID, ZooBank is effectively safeguarding against the possibility that LSIDs will not persist as an adopted style of GUIDs in biodiversity informatics. With an object identifier that is itself globally unique, ZooBank can strip the LSID “wrapper” (i.e., the first four components of the LSID) and still maintain a true GUID for its data objects. This GUID (the UUID) can then be “wrapped” in other existing (or not-yet-invented) resolution protocols and syntaxes, and thus stand the greatest chance of persistence over the long-term.

### *Defining “Registration”*

Somewhat surprisingly, one aspect of ZooBank that has not yet been well-defined is exactly what constitutes “registration”. Until the ICZN adopts procedures for electronic publication of new names and nomenclatural acts, the information that will ultimately end up in ZooBank begins (in virtually all cases) as ink on paper. Before this information can be entered into ZooBank, it must first be converted to electronic textual form. Whether this is accomplished by manual keystrokes, or by Optical Character Recognition (OCR) of electronically scanned pages, the information itself usually ends up in a computer database. Many, many such databases exist and predate the existence of ZooBank, and thus the majority of existing electronic content that will ultimately reside within ZooBank is not yet “registered”.

The first step in getting this existing electronic information into ZooBank is to import it into the ZooBank database. Because of the way the ZooBank database is set up, at the time of data import, every entry is automatically assigned its permanent UUID. However, this UUID is not converted into a formal ZooBank LSID until it has undergone some “pre-screening” process. Exactly what this process entails is one of the many policy details of ZooBank implementation that has not yet been formalized. However, some of the basic criteria for assigning a formal ZooBank LSID would likely include:

- Assurance that the same data object has not already been assigned a ZooBank LSID

- Assurance that the data object is indeed among the domains of objects included within ZooBank (i.e, a Nomenclatural Act, a Publication, and Author, or a Type Specimen)
- Assurance that the data object falls within the scope of ZooBank (e.g., that it is, in fact, relating to zoological nomenclature, as opposed to botanical nomenclature)
- Some minimum level of information content (i.e., minimum required data fields, which has yet to be defined).

Other, as yet undetermined criteria may also need to be fulfilled before the assignment of a ZooBank LSID.

The assignment of a ZooBank LSID is only the first formal step in the registration process. At least one (but possibly more than one) additional step is required to ensure that the information is complete and accurate. “Complete” can be defined in several ways, such as minimum required data fields, as well as minimum standards for what must actually be entered into those fields. “Accurate” is a somewhat nebulous term in this context, but is generally considered to mean that a qualified person (i.e., a person with substantial familiarity with the ICZN *Code*) has examined a copy of an original publication (or, perhaps, a facsimile of an original publication, such as high-resolution digital scans of the pages), and has verified that the information was accurately transcribed from the printed page to the ZooBank database, and that the original printed page meets the criteria for availability as prescribed by the ICZN *Code*. In this, we aim to adhere to the standards promoted by Sherborn.

Still unresolved is where, exactly, within this workflow of information capture, the word “registered” is appropriately applied. Furthermore, the specific steps of this workflow, who should be fulfilling those steps, and the exact process by which those steps are fulfilled, all remain among the details of ZooBank yet to be established through formal policy. Nevertheless, it seems reasonably safe to presume that the ZooBank data workflow will involve some sort of “staging” area for content prior to assignment of LSIDs; followed by a “validation” process whereby LSID-assigned data objects are scrutinized for completeness and accuracy; and finally some sort of indication that the record has been validated. In fact, we expect that by virtue of ZooBank being updatable by experts, the effect of having ‘many eyes’ examining the data will quickly bring it to a higher standard than Sherborn could achieve as a single individual scanning tens of thousands of entries over many years.

### *What Data Objects Should be Registered?*

There are two dimensions of how the scope of ZooBank data content needs to be defined. The first is in terms of what “domains” of objects should receive ZooBank LSIDs, and the second concerns the scope of instances within each domain.

Currently, four “domains” of data objects have been identified as falling within ZooBank’s purview: Nomenclatural Acts, Publications, Authors, and Type Specimens. Each of these has some direct bearing on Articles of the ICZN *Code*, and therefore represent logical objects to register in ZooBank.

A Nomenclatural Act exists in the form of a taxonomic treatment appearing within a publication deemed available by the ICZN *Code*. In the context of ZooBank, “Nomenclatural Acts” are those particular taxonomic treatments (or “usage instances”) that are themselves governed by the ICZN *Code*. Published treatments of new taxon names in the species-group, genus-group, and family group all constitute Nomenclatural Acts. Other kinds of Nomenclatural Acts include leptotypifications and neotypifications, emendations, and actions by first revisers. Other kinds of Nomenclatural Acts may or may not fall within the scope ZooBank. For example, the botanical *Code* governs acts that form novel combinations of pre-existing species group names within pre-existing or newly-created genus-group names. The zoological *Code* does not govern such acts *per se*, but they do have some relevance to the ICZN *Code* in that they may create secondary homonyms. Also, the *Code* does include provisions that apply to names above the family-group rank, but it is not immediately clear whether such names (and the Nomenclatural Acts that established them) should be included within ZooBank.

While it is clear that names and other Nomenclatural Acts that adhere to the provisions of the *Code* should be registered in ZooBank, less clear is whether names and Acts that fail to meet such provisions should also be registered. Including them would open a ‘Pandora’s Box of issues relating to the near-infinite scope of names for animals that are not governed by the *Code*. However, exclusion of ungoverned names could limit the practical utility of ZooBank, in that quick access to a list of names and acts unavailable under the *Code* would clearly provide a valuable function to the taxonomic community (Pyle & Michel, 2008). Balancing the practicality of initiating a useful list of known unavailable names with the Sisyphean task of assembling them needs to be weighed up in the development of ZooBank.

Regardless of what the final scope of taxonomic treatments that are defined as Nomenclatural Acts within ZooBank, all such Acts come to exist through Publications. Article 8 of the ICZN *Code* defines published works in the context of zoological nomenclature, and establishes criteria for determining whether such works are available under the *Code*. Certainly, all Publications that contain registered Nomenclatural Acts (however they are defined) should be included within the scope of registered ZooBank publications. But there may be other publications worthy of registration, such as works that have been explicitly rejected – if for no other reason than to make such explicit rejection clear to all ZooBank users (Pyle & Michel, 2008)..

Although not explicitly governed by the ICZN *Code* as Nomenclatural Acts or Publications, Authors of taxon names and Acts are integral to zoological nomenclature, and are addressed by several Articles in the *Code*. Again, it is clear that Authors of registered published works, would themselves be registered; but there are other kinds of Authors that may fall within the scope of ZooBank. For example, contributors to the ZooBank registry may also be registered as “Authors”, even if they have not authored registered publications.

The fourth domain within ZooBank – the “primary” or “name-bearing” type specimens (such as holotypes, syntypes, lectotypes, and neotypes) – obviously has important relevance to *Code*-governed zoological nomenclature. However, specimen data traditionally fall within the purview of the institutions that house natural history collections. The consensus among Commissioners at the Paris ICZN meeting in August 2008 was that type specimens represent a

legitimate domain within ZooBank, but it remains unresolved whether “secondary” (non-name-bearing) types may also be entered into the ZooBank registry.

A fifth domain of data object considered for inclusion within ZooBank by the ICZN Commissioners is the type-specimen repository. There are certain Articles in the *Code* that address the collections in which type specimens may be deposited, so there are grounds for including this domain as part of the ZooBank initiative. However, there are many unresolved details about how such entities would be defined, and to what extent their registration would, or would not have any bearing on the availability of names established based on type specimens housed therein. Moreover, at least two other initiatives (the Biodiversity Collections Index, and the Registry of Biological Repositories) already exist to serve this function, and it is not yet clear how ZooBank would interact with these other two registries.

### *Prospective Registration*

As originally conceived, the primary function of ZooBank is to serve as a registry for new Publications and Nomenclatural Acts, such that registration occurs before or immediately after the publication of a work containing Nomenclatural Acts (Polaszek 2005a, 2005b). This aspect of ZooBank serves the community by allowing the rapid dissemination of new information, and establishes a foundation upon which a future version of the ICZN *Code* would require that all new Publications and Acts be registered in order to be regarded as available under the *Code*.

The major obstacle to implementing “prospective” registration in ZooBank is in getting taxonomists to willingly contribute their time in order to enter the content into the ZooBank database. Compulsory registration, as enforced by the *Code*, would only be successful with a willing taxonomic community, so requiring registration will not, by itself, guarantee the success of ZooBank (i.e., it may instead lead to large-scale non-compliance with that requirement of the *Code*). The first step in garnering willing support from the taxonomic community is to make the processes of entering and editing records in ZooBank very simple and intuitive. Efforts in this regard are currently in development. But beyond merely lowering the “cost” of entering content by simple user interfaces, ZooBank needs to provide taxonomists with value-added features that help taxonomists get their jobs done. Some of these features could be implemented within ZooBank – such as free “subscription” services that notify members of the taxonomic community of new publications with nomenclatural content, perhaps increasing awareness of an author’s publication and subsequently increasing its citation value. But most of the benefits that ZooBank will be able to offer the community will likely involve integration with external data resources, which can help taxonomists locate and access information of direct relevance to their work. Steps are currently being taken to ensure ZooBank is tightly integrated with other informatics initiatives, but this is an area of ZooBank that needs further input from the community.

Another way to facilitate prospective registration is through cooperation with active journals that publish works containing information of nomenclatural importance. Two such journals (*Zootaxa* and *ZooKeys*) are actively engaged in the development of ZooBank, and have played an important role in keeping ZooBank moving forward. Other journals have expressed a strong interest in becoming similarly involved. Perhaps within the near-term future, a tight

alliance with ZooBank will increase the value of publications, enabling such journals to draw more quality manuscript submissions and enhance their impact factor.

We feel that the most challenging difficulty in establishing a regimen of prospective registration is in establishing the right policies and protocols of ZooBank. In broad strokes, some of the core principles are relatively easy to define. But as alluded to above, the difficulty lies in the specific details. Three general scenarios have been described for the general function and workflow for prospective registration (Polaszek et al. 2008; Pyle & Michel 2008). The only way to arrive at satisfactory answers to many of the outstanding questions is through maintaining active dialog among practicing taxonomists. Some of this dialog has already begun, but much more is still needed.

### *Retrospective Registration*

A separate, but equally important aspect of ZooBank is retrospective registration. This shares with prospective registration many of the detailed issues relating to minimum data requirements and validation. The task of populating ZooBank with complete legacy information on nomenclatural acts is not trivial. It should be kept in mind that there are an estimated 16,000 – 24,000 new additions for animal names yearly (N. Robinson, Zoological Record pers. comm., P. Bouchet, pers. comm.) to an estimated 1.7-1.8 million described animal species (Bouchet, 2006). As each species may have from one to ten (or even more) synonyms, the numbers of names to be checked for homonymy and primary synonymy is enormous. Estimates are often based on a few well-studied groups, but it should be underscored that the pattern of taxonomic work differs greatly among taxonomic groups.

There are a wide range of possible sources of such content, ranging from scanned literature with text derived from Optical Character Recognition (OCR) software, such as the increasing body of content originating from the Biodiversity Heritage Library, to citizen scientist initiatives as have worked for GalaxyZoo in astronomy ([www.galaxyzoo.org](http://www.galaxyzoo.org)) or Herbaria@home in collections management (<http://herbariaunited.org/atHome/>), to highly robust nomenclatural databases for certain groups of animals, which could serve as seed content for retrospective registration content. Bringing together information from disparate sources is a great challenge for a project like ZooBank that aims ultimately to have only information of the highest reliability. However, in the long term, successful fusion of information will allow ZooBank to exceed the qualities of its predecessors, the published nomenclators of Sherborn, Neave and von Schulze. Sherborn pointed to taxonomic catalogues as sources for the most informed data on scientific names. He commented, with a heartrending plea, that:

“Although much time has been expended in trying to secure the endless combinations and permutations of specific names, it is felt to be impossible for one human being to attain completeness in this direction by reason of the colossal amount of literature to be dealt with. Those who wish to gain this desirable result are referred to such works as the British Museum Catalogues of Birds, Marsupials, and Fossil Fishes, Brady's Report on the Foraminifera of the Challenger, Della Torre's Hymenoptera, Bronn's Index Palaeontologicus, Stiles and Hassall's Indexes to Worms, etc., where such attempts have been carried to successful conclusion. Still, a great mass of references has been here included

which it is hoped will have secured all generic and trivial names, and put the searcher on the track of a more complete synonymy. But even in this direction the methods adopted by many authors are such as to baffle the ingenuity of the recorder unless he happens to be a specialist in each group. Objection may be raised to those cases where the trivial name is referred back to a previous genus without a reference being given. I plead for compassion. Many hundreds of these cases have been pursued only to find that the author has quoted a previous author wrongly either by name or for the genus, and the time has been wasted. I am no longer young and, regrettable though it may be to me to leave such references unverified, I know my life is limited and I must press on.” (Sherborn 1921, p. ix).

ZooBank will be able to harness the expertise of many sources, without driving any single zoologist to a sorrowful end.

Once again, while the broad concept seems straightforward, the detailed implementation raises many issues. For example, what motivation would the owners and managers of such existing database content have in allowing portions of their databases – often having been built over decades of difficult effort – to be wholesale included within ZooBank? Further to this, how can ZooBank ensure substantial cross-linking back to original source content such as type collections, both in terms of providing appropriate credit for the data source, as well as enabling users of ZooBank to quickly access more detailed information that may be available on the original source? In the case of willing content providers, what standards of data quality should be applied to the imported content, and at what stage of validation could content be directly imported? How could such standards be applied across disparate data sets, which may have come about as a result of different priorities? For example, many nomenclatural databases include their own internal measures of quality and completeness; there would need to be protocols for translating these myriad quality metrics to the established ZooBank standards of quality (whatever those turn out to be).

The issues and associated questions are not trivial, and will likely only be resolved and answered on a case-by-case basis, following guidelines established collaboratively by the ICZN, the existing content holders, and the broader taxonomic community.

### *Quantity vs. Quality*

At the heart of many of these outstanding issues lies a fundamental interplay between data quantity, and data quality. Though often times represented as a trade-off, these two measures of ZooBank source content are not necessarily mutually exclusive. As mentioned previously, there may be several levels of validation along the entire process of ZooBank registration. That larger quantities of content may be imported at a low level of validation, does not preclude the establishment of rigorous policies and protocols for designating high levels of validation. As long as the current stage or level of validation is made clear on the ZooBank website, both quality and quantity can co-exist within the same system.

### **The Next 250 Years**

Though there is a persistent sense of urgency for ZooBank to quickly become adopted as the official online registry for zoological nomenclature, as mandated through provisions in a future version of the ICZN *Code*, there is an equal (if not greater) need to “get it right”. The Linnaean Enterprise has persisted for two and a half centuries, and the ICZN *Code* is itself more than a century old. While information technology and shifting paradigms of scientific discourse advance at an amazing pace, ZooBank must strike a careful balance between satisfying immediate needs, while both honoring a robust historical legacy and anticipating needs for the next 250 years.

Several initiatives are currently in the early stages of development, which will certainly have direct bearing on the role and function of ZooBank. The Global Names Architecture (GNA) – an effort to establish a common set of indexes and web services intended to cross-link taxonomic data from a broad array of sources, will rely heavily on nomenclatural authorities such as ZooBank and its associated content providers to build such cross-links. Large initiatives such as the Global Biodiversity Information Facility (GBIF) and the Encyclopedia of Life (EoL) have revitalized the perceived need to find harmony among the different nomenclatural *Codes*.

There is no doubt that, as the ICZN moves forward with such initiatives as the acceptance of electronic forms of published works (ICZN, 2008) and a revised 5<sup>th</sup> Edition of the *Code*, and the internet fundamentally transforms the way scientific information is communicated among researchers, the Linnaean Enterprise is on the cusp of a fundamental paradigm shift. Whether or not it maintains the same level of relevance it has enjoyed these past two and a half centuries during the 250 years to come, may well depend on how carefully the taxonomic community of today navigates the uncharted waters in now finds itself in.

## Acknowledgements

We wish to thank Andrew Polaszek for organizing a very stimulating symposium, Jon Todd, Ken Johnson, and the many ICZN Commissioners and other interested taxonomists, especially on the ICZN listserv, who have contributed many of the ideas included herein. Keri Thompson (Smithsonian Libraries) and David Remsen (GBIF) provided current usage statistics for *Index Animalium* and *Nomenclator Zoologicus*.

## References

- Agassiz, L. (1848) *Nomenclatoris zoologici index universalis : continens nomina systematica classium, ordinum, familiarum et generum animalium omnium tam viventium quam fossilium... homonymiis plantarum* (1848). Soloduri : Sumptibus Jent et Gassman. 1135 pp.
- Bouchet, P. (2006) The magnitude of marine biodiversity. In: Duarte, C. (Ed.) (2006). *The exploration of marine biodiversity: scientific and technological challenges*. pp. 31-62.
- ICZN (2008) Proposed amendment of the *International Code of Zoological Nomenclature* to expand and refine methods of publication. *Zootaxa*, 1908: 57–67.
- Linnaeus, C. (1758) *Systema naturae per regna tria naturae, secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis*. Tomus I. Editio decima, reformata. 10th Ed., Vol. 1, pt. 1. Holmiae, 1, ii+824 pp.

- Marschall, A.F. (1873) *Nomenclator zoologicus continens nomina systematica generum animalium tam viventium quam fossilium, secundum ordinem alphabeticum disposita, aspiciis et sumptibus C.R. Societatis zoologico-botanicae / conscriptus a comite. Ueberreuter (M. Salzer)*, 482pp.
- Melville, R.V. (1995) *Towards Stability in the Names of Animals: A History of the International Commission on Zoological Nomenclature 1895-1995*. International Trust for Zoological Nomenclature, London. 104 pp.
- 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. (also available online <http://uio.mbl.edu/NomenclatorZoologicus/browse.html>)
- Polaszek, A., Agosti, D., Alonso-Zarazaga, M., Beccaloni, G., de Place Bjørn, P., Bouchet, P., Brothers, D.J., Earl of Cranbrook, Evenhuis, N., Godfray, H.C.J., Johnson, N.F., Krell, F.-T., Lipscomb, D., Lyal, C.H.C., Mace, G.M., Mawatari, S., Miller, S.E., Minelli, A., Morris, S., Ng, P.K.L., Patterson, D.J., Pyle, R.L., Robinson, N., Rogo, L., Taverne, J., Thompson, F.C., van Tol, J., Wheeler, Q.D. & Wilson, E.O. (2005a) Commentary: A universal register for animal names. *Nature*, 437, 477. ([http://www.iczn.org/Nature\\_Commentary.pdf](http://www.iczn.org/Nature_Commentary.pdf))
- Polaszek, A., Alonso-Zarazaga, M., Bouchet, P., Brothers, D.J., Evenhuis, N., Krell, F.-T., Lyal, C.H.C., Minelli, A., Pyle, R.L., Robinson, N.J., Thompson, F.C. & van Tol, J. (2005b) ZooBank: the open-access register for zoological taxonomy: Technical Discussion Paper. *Bulletin of Zoological Nomenclature*, 62, 210–220. ([http://www.iczn.org/ZooBank\\_Paper.htm](http://www.iczn.org/ZooBank_Paper.htm))
- Polaszek, A., Pyle, R. & Yanega, D. (2008) Animal names for all: ICZN, ZooBank, and the New Taxonomy. pp. 129–142. In: Wheeler, Q.D. (Ed.). *The New Taxonomy*. CRC Press, Boca Raton. 237 pp.
- Pyle, R. (2008) Summary of session on ZooBank (session 5). *Bulletin of Zoological Nomenclature* 65: 257-260.
- Pyle, R.L., Earle, J.L. & Greene, B.D. (2008) Five new species of the damselfish genus *Chromis* (Perciformes: Labroidae: Pomacentridae) from deep coral reefs in the tropical western Pacific. *Zootaxa*. 1671, 3–31. (<http://www.mapress.com/zootaxa/2008/f/zt01671p031.pdf>)
- Pyle, R.L. & Michel, E. (2008) ZooBank: Developing a nomenclatural tool for unifying 250 years of biological information. In: Minelli, A., Bonato, L. & Fusco, G. (eds) *Updating the Linnean Heritage: Names as tools for thinking about animals and plants*. *Zootaxa*, 1950: 39–50.
- Remsen, D.P., Norton, C. & Patterson, D.J. (2006) Taxonomic Informatics Tools for the Electronic *Nomenclator Zoologicus*. *Biological Bulletin* 210: 18–24.
- Scudder, S.A. (1882) *Nomenclator zoologicus*. An alphabetical list of all generic names that have been employed by naturalists for recent and fossil animals from the earliest times to the close of the year 1879. *Bulletin of the United States National Museum* 19. 340pp.
- Sherborn, C.D. (1902-1933) *Index animalium; sive, Index nominum quae ab A.D. MDCCLVIII generibus et speciebus animalium imposita sunt*. Trustees of the British Museum. 7056 pp. (also available online <http://www.sil.si.edu/digitalcollections/indexanimalium/>)
- 
-