Taxonomic Informatics Tools for the Electronic *Nomenclator Zoologicus*

DAVID P. REMSEN,* CATHERINE NORTON, AND DAVID J. PATTERSON

Marine Biological Laboratory, Woods Hole, Massachusetts 02543

Abstract. Given the current trends, it seems inevitable that all biological documents will eventually exist in a digital format and be distributed across the internet. New network services and tools need to be developed to increase retrieval rates for documents and to refine data recovery. Biological data have traditionally been well managed using taxonomic principles. As part of a larger initiative to build an array of names-based network services that emulate taxonomic principles for managing biological information, we undertook the digitization of a major taxonomic reference text, *Nomenclator Zoologicus*. The process involved replicating the text to a high level of fidelity, parsing the content for inclusion within a database, developing tools to enable expert input into the product, and integrating the metadata and factual content within taxonomic network services. The result is a high-quality and freely available web application (http://uio.mbl.edu/NomenclatorZoologicus/) capable of being exploited in an array of biological informatics services.

Introduction

In 1969, Raath described a new species of dinosaur with fused foot bones that he named "*Syntarsus*," (Raath, 1969), unaware that a hundred years earlier a beetle had been named *Syntarsus* by Fairmaire (Fairmaire, 1869). The codes of nomenclature explicitly disallow the use of the same name for two organisms. The mistake was discovered by CSIRO's Adam Slipinski, a beetle expert who started a dispute over nomenclatural ethics (Holden, 2002) when he and his colleagues proposed a replacement name, *Megapnosaurus*, which roughly translates to "big dead lizard" (Ivie *et al.,* 2001).

The significance of this tale is that the creation of the duplicate name, a homonym, did not come to light until 2001, over 30 years after the name was used a second time. During this period, the word *Syntarsus* lacked an unambiguous meaning. This reflects the poor state of information management for biology (Agosti and Johnson, 2002; Stein, 2002). A systematic process for naming organisms has been in place for over 250 years and in the case of animals is regulated by the International Code of Zoological Nomenclature (International Commission on Zoological Nomenclature, 1999). A key task is to assign unique and formalized names to organisms. In the age of digitization and gigabyte data systems it may come as a bit of a surprise that a unified and comprehensive catalog of names used for living (or once-living) organisms does not exist. Efforts to create a comprehensive online compendium of code-compliant animal names are only now starting (Patterson, 2003; Patterson *et al.,* 2003; Polaszek *et al.,* 2005; Thorne, 2003).

The catalog of all of the estimated 1.75 million species (Wilson, 2003) that have been described would be represented by a list of *current* valid code-compliant names. For informatics purposes, we need to compile all names that have ever been used to refer to taxa. A catalog of recorded names of living and extinct taxa will be substantially larger and more encompassing than a compilation of code-compliant names. Each species may be represented by two or even dozens of previously valid names, as well as by an array of lexical variants, mistypings, and vernacular names. Such names, valid or invalid, spelled correctly or not, annotate data relating to organisms. Together, they form an extensive vocabulary of metadata terms that can be exploited for data search and retrieval. That role can be enhanced by supplementary ontologies which link together names that refer to the same organisms or which place the names within hierarchical arrays. Such extensions underpin taxonomic indexing services (Patterson *et al.,* 2006) that can

overcome challenges in finding information for organisms whose names have changed, because they help to determine if a single name has been used to refer to more than one organism or if there is more than one name for a taxon, and they can provide a general taxonomic placement for a name.

The Universal Biological Indexer and Organizer (uBio) project was established at the Marine Biological Laboratory Woods Hole Oceanographic Institution (MBL/WHOI) Library in response to the need for a comprehensive compilation of names and their relationships. With support from the Andrew W. Mellon Foundation, the MBL/WHOI library has developed a suite of network tools that revolve around a central Taxonomic Name Server (www.ubio.org). In 2004, the MBL/WHOI Library identified a number of taxonomic texts as priority targets for digital conversion. These texts were prioritized because of their nomenclatural coverage or because they allowed the exploration of modeling taxon concepts. They included a Smithsonian taxonomic bulletin, *The Catalog of Living Whales* (Hershkovitz, 1966), and now available at http://uio.mbl.edu/Hershkovitz/; and *Nomenclator Zoologicus* (Neave, 1939–1996).

A key component of the uBio strategy for assembling a compilation of names has been to catalog names of genera. A name that is given to a species is in the form of a binomial (*Syntarsus kayentakatae*) with the species name preceded by a parent genus. As there are about 10 species per genus on average, and as determining the identities of genera is considerably easier than determining the identities of species, a compilation of all generic names would require two or three orders of magnitude less effort than cataloging all species names (Patterson, 2003). A compilation of generic names provides a dictionary that can be used in the automated discovery of species names in documents, and also provides a framework around which species names can be assembled. The compilation of generic names therefore allows for the more rapid introduction of a taxonomic cyberinfrastructure for all taxa, and will accelerate the compilation of all names of all species.

Nomenclator Zoologicus is a catalog of the bibliographic origins of the names of every genus and subgenus in the published literature since the tenth edition of Linnaeus' *System Natureae* in 1758 (Linnæus, 1758) up to 1994. An estimated 340,000 genera are represented in the text and there are approximately 3000 supplemental corrections. It provides a nucleus of core genera data and is recognized as an essential reference document by the zoological taxonomic community. The list provides bibliographic details to allow the original descriptions to be found, and provides synonymies and general taxonomic placement for useful information retrieval purposes. Moving *Nomenclator Zoologicus* from a print to a web database interface creates opportunities for new tools and enhances inquiry. Search queries cross all volumes instantly. Hundreds of thousands of records can be collated and summarized to reveal patterns that would be completely impractical to compile any other way. A quick search of the database (Fig. 1) reveals the *Syntarsus* problem.

Methods: Producing the Digital Document

The MBL/WHOI library worked in close collaboration with the Zoological Society of London, the publisher of the *Nomenclator Zoologicus*.

Names of an estimated 340,000 genera (Table 1) are listed in *Nomenclator Zoologicus* alphabetically. Each has a bibliographic reference to the original description and an indication of the animal group to which it belongs. There are approximately 3000 supplemental corrections.

The bibliographic records follow a relatively consistent format containing the name, author reference, year of publication, publication reference, and a general taxonomic category for the genus (Fig. 2). In addition, some records contain nomenclatural or cross-reference annotations. A dagger (†) indicates an extinct taxon.

The common components of the records were used as a framework for parsing the records into a set of columns (Table 2) as a prelude to moving the contents into a database management system.

In respect to conversion of the text to a digital format, pages were first scanned manually, and the resultant image files were passed to a commercial optical character recognition application that resulted in text conversion accuracies ranging from 95%–99%. This accuracy rate was unacceptably low because it required excessive pre-release editorial

Figure 1. The homonyms—*Syntarsus* in the digital *Nomenclator Zoologicus*.

Table 1

Composition of the Nomenclator Zoologicus *volumes—including individual corrigenda and addenda*

Volume	Dates	Pages	Records*
1 [A-C]	1758-1935	957	53,944
2 [D-L]	1758-1935	1025	58,007
3 [M-P]	1758-1935	1065	60,447
4 [Q-Z]	1758-1935	758	42,562
5	1936-45	308	18,310
6	1946-55	329	18,556
7	1956-65	374	20,249
8	1966-77	620	29,703
9	1978-94	747	41,146
Total		**6183**	**343,143**

verification, so the approach was rejected in favor of using a commercial text conversion company offering 99.995% accuracy.

The converted files were provided as UTF-8 encoded, tab-delimited text files corresponding to the individual volumes. In addition to the seven columns identified from the actual text (name, author, year, publication, group, extinct, annotation), additional fields were added to indicate the source volume and page number for each record.

The text files were then imported into a desktop database management system, (Filemaker Pro 7.0) for an initial round of quality assessment. A number of quality tests were run to evaluate the quality of the conversion process. Material was re-digitized if it failed to achieve high quality.

One test examined columns known to contain a particular class of data and searched for exceptions. Page and volume columns, for example, should contain only integers. Simply sorting the columns allowed all non-integer values to be grouped together for scrutiny. A second approach was to export a summarized list of distinct column values. The group column is expected to contain zoological group names only. There were fewer than 3500 unique entries in this field for the entire 340,000+ records. Within such a short list, erroneous data such as integers, authors, or publication information are easily identified. Other tests involved searching for blank records where data should appear, or locating particular terms such as "See"—a common component in the Annotation field (*e.g.*, "See *Actaeonema* Conrad 1865). The occurrence of strings such as this within other columns revealed parsing errors.

Patterns assisted in the parsing of the converted data into columns. The 'Group' field is formed by a name preceded by a dash and is often the last element in a record. Using an expression "a dash followed by a word represents the end of a record" holds true in the majority of cases, but in some cases a dash was a legitimate part of a different column, such as within a publication reference. In these instances, the record would be prematurely truncated. As a consequence, future conversions of similar documents would benefit from having two versions of the converted file available for review—the final parsed version and an unparsed raw form. The lengths of corresponding record pairs could be compared, and these would reveal any cases of truncation. This is desirable because, after the final editorial rounds, truncations are the main source of editorial corrections.

An array of techniques were employed to locate and identify typographical errors. Searches within the authority year column are expected to find dates beginning with 18** and 19**. Searches were made for strings containing "i8," "i9," "l8," or "l9," where a numeric 1 was mistakenly interpreted as the letters "i" or "l." Other optical character reading errors included the conversion of the name "Brünn" to "Briinn." We manually checked pages where the names contained diacritical marks. Such errors were sufficiently frequent that they required five iterations, and the final process involved "double-keying." This confirmed the view that optical character reading methods were inadequate to meet the challenges of creating a high fidelity electron version of the text of *Nomenclator Zoologicus*.

Once vetted to adequate standards, the converted volume files were imported to mySQL. The contents of all nine volumes were collated into a single table and assigned unique sequential record identifiers. Several additional columns were added at this stage. The "Corrigenda Flag" identifies records that are part of the Addenda and Corrigenda sections of volumes 4–9. Corrigenda records include a second reference to a name, and the flag allows them to be discriminated from true homonyms. The attribute was set by

Bathysphaera Beebe 1932, Bull. New York zool. Soc., 35, 175. – Pisces

Cylindrus (emend. pro *-der* Montfort 1810) Deshayes 1824, Dict. Class. H.N., **5**, 236. – Moll.

Bathystoma Marsson 1887, Pal. Abh., 4, 88.– Bry. (See Bathystomella Strand 1928.)

Figure 2.   Several typical records from *Nomenclator Zoologicus*.

Table 2

Records as parsed into columns

Name	Author	Year	Publication	Group	Extinct	Annotation
Bathysphaera	Beebe	1932	Bull. New York zool. Soc., 35, 175	Pisces	No	
Cylindrus	Deshayes	1824	Dict. Class. H.N., **5**, 236	Moll.	No	(emend. pro - *der* Montfort 1810)
Bathystoma	Marsson	1887	Pal. Abh., 4, 88	Bry.	No	(See Bathystomella Strand 1928.)

applying the value of '1' to all records falling inside the Corrigenda page range for each volume. All other records received a value of 0. An addenda flag represented new records (not duplicates) within the Addenda and Corrigenda. A "homonym flag" was set for all records that included a string in the "name" column that was duplicated in any other record for any reason. This flag was applied to all true homonyms and duplicate records, and it served as an alert that the record may require further scrutiny.

Approximately 61,000 records contain information within the "annotation" column. Of these, about 55,000 refer to different names within the collection. These cross-references usually identify a synonym or orthographic variant of the name, such as **'*Abala* (err. pro Ababa Casey 1897)'**. In a significant number of cases, the cross-referenced name was incomplete [**'*Abanchogaster* (pro-gastra Perkins 1902)'**] and required intervention to infer the actual name—which in this case is *Abanchogastra*. The cross-references were mapped in stages, starting with automated processes and proceeding to manual review as required. A combination of custom perl and PHP scripts were employed with databased components to assist in the process.

Results: The Product and Editorial Applications

The online version of *Nomenclator Zoologicus* is composed of a set of PHP scripts interfacing a mySQL database running on a Linux-based computer. The online application has three major components: the database interface, a page image browser, and supporting documentation. This is supplemented with an online editor.

The database interface is divided into three primary components: a search interface consisting of a simple and advanced search form; a search results interface providing paged, tabular output of query results; and a record detail page that contains the full data record, associated cross-reference information, and user-annotations.

The simple search feature provides a single primary input field that, by default, searches all the text-containing columns in the database using a "contains" search qualifier—similar to the popular online search engines. The search function allows some limits to be added to the query and allows searches for specific volumes or pages. An advanced search option provides input fields for all six string-containing columns for more precise Boolean searching. The "con-tains" qualifier can be turned off for more precise searching and file globbing operators (*e.g.,* "Ab*" to find strings beginning with 'Ab') are supported.

The search results page provides a paged tabular view of search results. The results are divided into page groups of 500 records with page navigation options both top and bottom. Each displayed record consists of the entire core data record. A set of icons preceding each record provides addition links and qualifiers. A hyperlink on the name string leads to a record detail page. (Table 3)

The record detail differs from the tabular record in the results page only by mapped cross-references or reviewer annotations (if present). The record page is the point where a user can add new annotations.

Digital page images are available in both PNG and PDF format. They can be accessed by the search interface and by a separate page browser. The browser interface is intentionally simple. Users begin page browsing *via* an image-mapped representation of the nine volumes on a bookshelf. The front matter from each volume is linked separately *via* numbered links. A previous and next button navigates through the pages, or a user can enter a volume and/or page number to jump to that page image. The data represented in a page is hyperlinked to the search results page.

The documentation of the online application contains background information on the project, technical details regarding the development of the database, a schema, and some pre-computed results of queries not available in the online application. These include record summaries

Table 3

Record icons and their meaning

	Review Status: record has not been manually verified
	Review Status: record has been anonymously manually verified
	Review Status: record has been manually verified
	Hyperlink to the PNG page image containing the record
	Record contains cross-reference information via the record detail page
	Record contains annotations submitted by one or more users

grouped by year and author, as well as a complete list of homonymous names. This format is not completely accurate because of duplicates within the volumes themselves, independent of the Corrigenda. Homonyms are identical names that refer to different taxa. Identification of homonyms within the *Nomenclator Zoologicus* was confounded by the occurrence of duplicate records within the text. The procedure for setting the homonym flag was to examine potential homonym groups. If the sole members of the group were determined to be identical, the flag was set to zero. If at least two members of a group were determined to be different, the flag was set to one. Members of these groups may still contain some duplicates, which are retained to preserve the fidelity of the original.

The online version of *Nomenclator Zoologicus* was announced and made public in December 2004 *via* the uBio website (http://www.ubio.org) where the work was undertaken, and an email announcement was sent to the email-based list server TAXACOM, a biological systematic and biocollections discussion list. The positive response to the *Nomenclator Zoologicus* online version led to the next steps for the data conversion. The high quality of the final draft from the contractor, combined with our automated and assisted review tools, assured that the released version was of a very high fidelity, but a manual review could bring the overall quality of the conversion to nearly 100%.

An online editorial application was developed to enable a wide community of experts in the taxonomic community to edit and annotate the electronic *Nomenclator Zoologicus* as a part of the process of quality control (Fig. 3). The application simplifies the task of comparing the new digital records with the original printed version.

Shortly after the release of the first online version, a mailing was sent to the TAXACOM list-serve seeking volunteers from the community to assist in a manual review of the records. Volunteers were assigned 100-page blocks and were asked to compare each digital record with its corresponding print original to ensure that the conversion introduced no new errors.

The application consists of a combination of PHP code and JavaScript. The application presents a screen containing both the page image and the converted digital record. The page image and the digital record can be positioned independently in order that the two can be aligned. When the two records are optimally aligned, it is relatively easy to compare the two records.

When a record is reviewed, the reviewer has three options. The first affirms that the two records match, and the next record is presented. The second "Correct" option provides a form where the reviewer can make a correction to bring the record in concordance with the original. In actuality this correction is made to a duplicate record that is kept separate until a further review determines whether the change is accurate. If it is, the correction is made. The third option allows the reviewer to add new annotations to the record. There are numerous, and in some cases, well-known errors within the *Nomenclator Zoologicus*. These errors are part of the printed record, and are preserved. These errors have not been corrected, but the records are annotated using a "Comment" option.

In the application, a JavaScript-based red horizontal rule can be placed on top of the page image to help locate items in the print record. After reviewing a record and proceeding to the next, the application can scroll the page image to the next record. This requires the page image to scroll by the correct amount. There is no direct correlation between the
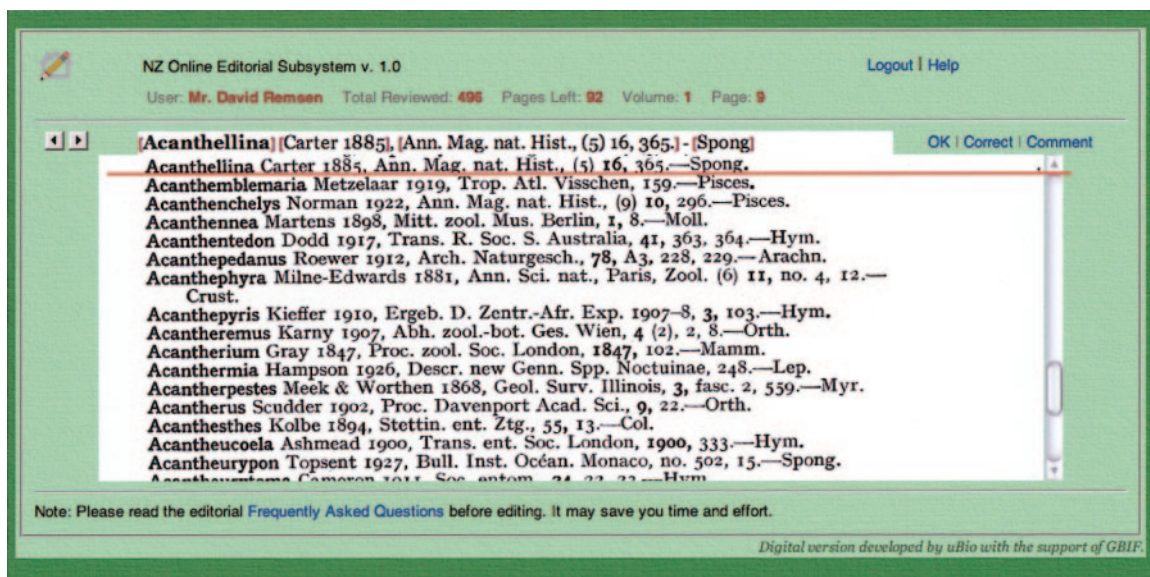

Figure 3. The *Nomenclator Zoologicus* editorial review application.

page image file (a PNG or PDF file) and the resultant digital record, so this is not a simple requirement. The application relies on knowledge that, on average, a line is composed of 119 characters and is 12 pixels in height. There remain some challenges because of imprecisely aligned images. This problem is corrected by allowing the reviewer to manually position the page at any time. The digital record can be moved horizontally to align the left boundaries of the two records for easier comparison.

During the past 6 months, expert taxonomists worldwide have reviewed 877,176 characters— tens of thousands of records—and verified the accuracy of the initial conversion. To date, only 33 characters have required correction, indicating that the digital conversion process achieved an accuracy rate of 99.97%.

Volume 10 of the printed version of *Nomenclator Zoologicus* was provided in digital format in October 2005 and was added without complications to the database.

Discussion

The primary uses of the electronic *Nomenclator Zoologicus* are the same as those of the printed version. It can be used to establish whether a name has already been used for a genus, and to locate the source of a code-compliant name. Since going online, 232,568 searches have been performed on the collection of names. Only 52,761 pages (full-screen shots) were browsed, indicating that most users rely on the converted data and do not verify from the page image. Queries have come from 5503 unique IP addresses. Although this may appear trivial compared to most web statistics, it is worth bearing in mind that the taxonomic community has been estimated as having as few as 6000 individuals (Wilson, 2003). The usage statistics of the online *Nomenclator Zoologicus* suggest that this is an underestimate.

Nomenclatural compilations are invaluable to avoid the creation of homonyms. A simple search in the online *Nomenclator Zoologicus* version identifies more than 21,000 homonym groups, with some of the most common generic homonyms listed in Table 4. The availability of such tools

Table 4

The most commonly used homonyms and their frequency of use

Wagneria	13
Arqus	11
Carinella	11
Melia	11
Acanthonotus	10
Discus	10
Hoplites	10
Nicholsonia	10
Pandora	10
Trachynotus	10

would have solved the *Syntarsus* problem in an instant.

Certain useful information summaries, such as those included in the online documentation section, are relatively easy to generate from a digitized version of the data. A summary by author, for example, reveals that Linnaeus was not even in the top 100 most referenced authors (he was 123rd).

In response to a request, we have examined the suffixes of genera names for evidence in favor of developing standard conventions for suffixes of generic names. It might be realistic to use standardized endings, such as the –idae ending for families of animals and –ini for tribes. Ninety-one percent of all genera names in *Nomenclator Zoologicus* end with -a, -s, or -m. This insight has proven valuable in other contexts. As indicated earlier, the uBio project has developed tools to discover names in source documents. Our compilation of generic names forms a dictionary that helps to confirm that a string refers to a species name. Knowing the most likely termini of name-strings is also used in our names recognition tools.

The development of the online *Nomenclator Zoologicus* is a significant step toward meeting the informatics needs of taxonomists and in providing the foundations of informatics tools for biological information management. The online version of *Nomenclator Zoologicus* will remain a stand-alone web site, but it is also currently being incorporated into the NameBank names registry that already holds almost 4 million name strings. The enhancements include the cataloging of genera that are in NameBank but were not in the original *Nomenclator Zoologicus*. This will allow the original to remain distinct yet also a component of this larger collection and will make the names accessible *via* web services for more flexible and widespread use.

The inclusion of the zoological genera missing from *Nomenclator Zoologicus*, coupled with lists of genera of plants, fungi, prokaryotes, and protists, is providing the foundation for the accelerated assembly of a compendium of all names of all species. That compendium serves as the foundation layer of a multi-part biological names-based cyberinfrastructure for biology.

As the use of *Nomenclator Zoologicus* online continues to grow, taxonomists have offered additional lists of names to supplement the collection. This response reflects the value of a unified and comprehensive listing and the rewards of the internationalization of taxonomy through a cyberinfrastructure.

Literature Cited

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

Fairmaire, L. 1869.   Notes sur les Coléoptères recueillis par Charles Coquerel à Madagascar et sur les côtes d'Afrique. 2e Partie. *Annales de la Société Entomologique de France, 4 Série* **9:** 179–260.

Hershkovitz, P. l966.   *Catalog of Living Whales.* U.S. National Museum Bulletin No. 246. Smithsonian Institution, Washington, DC.

Holden, C. 2002.   Taxonomic tussle. *Science* **295:** 1459.

International Commission on Zoological Nomenclature. 1999.   *International Code of Zoological Nomenclature,* 4th ed. International Trust for Zoological Nomenclature, c/o Natural History Museum, London.

Ivie, M. A., S. A. Slipinski, and P.Wegrzynowicz. 2001.   Generic homonyms in the Colydiinae (Coleoptera: Zopheridae). *Insecta Mundi* **15:** 63–64.

Linnæus, C. 1758.   *Systema Naturae per Regna Tria Naturæ: Secundum Classes, Ordines, Genera, Species, cum Characteribus, Differentiis, Synonymis, Locis,* 10th ed. Laurentii Salvii, Stockholm, Sweden.

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.

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

Patterson, D.J., D. Remsen, and C. Norton. 2003.   Comment on Zoological Record and registration of new names in zoology. *Bull. Zool. Nomencl.* **60:** 297–299.

Patterson, D. J., D. Remsen, W. A. Marino, and C. Norton. 2006.   Taxonomic indexing—extending the role of taxonomy. *Syst. Biol.* (in press).

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

Raath, M.A. 1969.   A new coelurosaurian dinosaur from the Forest Sandstone of Rhodesia. *Arnoldia* **4:** 1–25.

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

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

Wilson, E. O. 2003.   The encyclopedia of life. *Trends Ecol. Evol.* **18:** 77–80.