IRMNG 2006–2016: 10 YEARS OF A GLOBAL TAXONOMIC DATABASE

TONY REES[1][1], LEEN VANDEPITTE[2], WIM DECOCK[2] AND BART VANHOORNE[2]

[1]*Private address, New South Wales,* Australia. [2]*Flanders Marine Institute/Vlaams Instituut voor de Zee, Wandelaarkaai 7, 8400 Oostende, Belgium*

Abstract.—IRMNG, the Interim Register of Marine and Nonmarine Genera, was commenced in 2006 as an initiative of the Australian OBIS Node (OBIS Australia) following an analysis of the taxonomic names management needs of the Ocean Biogeographic Information System (OBIS). The main objectives were to produce a hierarchical classification of all life, both extant and fossil, to at least generic level (and to species as data were readily available) and to provide a tool to distinguish marine from nonmarine, and extant from fossil taxa. Over its first 10 years of operation IRMNG has acquired almost 487,000 of an estimated 510,000 published genus names (including both valid names and synonyms) in addition to almost 1.8 million species names, of which 1.3 million are considered valid. Throughout this time IRMNG data have been available for public query via a dedicated web interface based at CSIRO in Australia, as well as being supplied as bulk downloads for use by a range of global biodiversity projects. Over the period 2014-2016 responsibility for the system has been passed to the Data Centre Division of the Flanders Marine Institute (VLIZ) in Belgium, which is continuing the maintenance and development of IRMNG at its new web location, www.irmng.org. With its present estimated holdings of >95% of all published genus names (plus associated authorities and years of publication) across all taxonomic domains, including fossil as well as extant taxa, within an internally consistent taxonomic hierarchy, IRMNG is at present uniquely placed to provide an overview of "all life" to at least generic level, to permit the discovery of trends in publication of genera through time, to provide preliminary information on the marine vs. nonmarine and extant vs. fossil status of the taxa concerned, and to generate lists of both unique and non-unique names (homonyms *sensu lato*) for the benefit of users of biodiversity data.

Key words.—Taxonomic databases; biodiversity; marine taxa; terrestrial taxa; extant taxa; fossil taxa; biological classification.

---

[1] Email address for correspondence: Tony.Rees@marinespecies.org.

INTRODUCTION

The desire to obtain an overview of taxon names for "all life", whether extant-only or also including named fossil taxa, is one which has repeatedly emerged since the time of Linnaeus in the eighteenth century, and has been revisited at intervals for various groups e.g., for animals (Agassiz 1848, Sherborn 1902-1933, Neave 1939-1996), plants (Hooker and Jackson 1895, Andrews 1970 plus supplements, Willis 1973, Farr et al. 1979), prokaryotes (Euzéby 1997-current), and viruses (International Committee on Taxonomy of Viruses 1971-current). Since the advent of the internet age a number of these works have been migrated to the web and/or new initiatives started, in particular the Integrated Taxonomic Information System ITIS)[2], the "Species 2000" collective[3], and their combined initiative the Catalogue of Life[4], currently in its 15th annual edition (2016). For extinct (fossil) taxa, the Paleobiology Database[5] is also becoming progressively more complete. Nevertheless, both of these latter two compilations still have some gaps: the 2016 Catalogue of Life currently covers only 84% of world diversity[6], while coverage of the Paleobiology Database is limited by its user contributions, and currently missing at least 100,000 valid fossil species names (data presented later in this paper).

On account of such gaps in coverage it is still possible to encounter both species names and those of genera, either within the scientific literature or included among names submitted to biological information systems as identifiers for accompanying data, not presently held in the above compilations and which cannot therefore be placed taxonomically, and/or ancillary information discovered, without additional manual effort.

Genesis of this work

Biodiversity data aggregation projects such as the Ocean Biogeographic Information System (OBIS[7]) and the Global Biodiversity Information Facility (GBIF[8]) commenced in the early 2000s with an aim to bring together available occurrence data on extant species in either just the marine domain (OBIS) or all habitats (GBIF). Incoming data to such projects typically comprise location data (latitude-longitude) together with an associated species name and potentially other information. OBIS and GBIF therefore require a taxonomic hierarchy in which to place the incoming data by means of the supplied taxonomic names, as well as (in OBIS' case) the capacity to discriminate marine from nonmarine taxa (and by extension, extant taxa from fossil) to support the intended purpose of displaying content for marine, extant species only.

In 2003-4 a process was commenced to extract all species names in the then-latest version of the Catalogue of Life (2003 edition), manually assign them a marine/nonmarine flag, and use this process to place names held in OBIS in the Catalogue of Life taxonomic hierarchy along with an indication of their marine/nonmarine status (see Rees and Zhang 2007). However, at that time a non-trivial proportion (around 30%) of OBIS names held could not be taxonomically resolved using the Catalogue of Life, prompting a search for an alternative method of name resolution together with a means to query their associated marine/nonmarine, and extant/fossil status.

The solution adopted was to attempt to create a more comprehensive index of genera only. This was considered more tractable for two reasons: first, the number of names to be compiled (valid names plus synonyms) is potentially an order of magnitude smaller for genera than for species, and second, a number of genus-level compilations were already in existence, which in combination could offer greater taxonomic completeness than available equivalents at specific level. For taxa covered by the zoological Code (International Commission on Zoological Nomenclature 1999) there exists a digitised online version of Neave's *Nomenclator Zoologicus*, updated to the end of 2004[9], for taxa covered by the botanical Code (McNeill et al. 2012) the online version of *Index Nominum Genericorum*[10], while prokaryotes are covered by the List of Prokaryotic Names with Standing in Nomenclature[11] and viruses by the taxonomy releases of the International Committee on Taxonomy of Viruses[12]. *Nomenclator Zoologicus*

[2] http://www.itis.gov/
[3] http://www.sp2000.org/
[4] http://www.catalogueoflife.org/
[5] https://paleobiodb.org/
[6] http://www.catalogueoflife.org/content/frequently-asked-questions#8
[7] http://www.iobis.org/
[8] http://www.gbif.org/

[9] http://ubio.org/NomenclatorZoologicus
[10] http://botany.si.edu/ing/
[11] http://www.bacterio.net/
[12] http://www.ictvonline.org/taxonomyReleases.asp

and *Index Nominum Genericorum* incorporate indications of extant/fossil (but not marine/nonmarine) status in most cases, plus family assignment is provided in the case of the botanical, prokaryote and virus name compilations. In practice, the project was started without relying on these particular resources, but they were made available within the first 2 years of the project (see Appendix 1).

The value of an all-genera list for indexing species data lies in the binomial format of species names in all domains except viruses, whereby the including genus is represented as the first element of the binomial name. Therefore, even if information on a particular species is not held explicitly, by inference it can frequently be presumed to inherit particular traits including taxonomic placement and marine vs. nonmarine, and extant vs. fossil information, to the extent that the latter two aspects are unambiguous (e.g., marine only, extant only). Further, the process of creation of relevant trait information within the database can be simplified via hierarchical propagation from higher to lower levels, (for example all echinoderms are marine, all trilobites are extinct), so that detailed manual effort is reserved only for "mixed cases", namely genera or higher taxa containing both extant and fossil, and/or both marine and nonmarine components. A supplementary benefit of creating and populating a taxonomic framework to genus-level is that the process of later adding species (when required) is simplified, since in the majority of cases their appropriate parent at generic level will already be present and can be used as an attachment point for the species concerned, with no additional effort needed to decide on their higher taxonomic placement.

The name applied to the project, the Interim Register of Marine and Nonmarine Genera (IRMNG), was intended to reflect that of ERMS, the European Register of Marine Species (Costello et al. 2001), a precursor to the World Register of Marine Species or WoRMS (WoRMS Editorial Board 2007-current), while indicating that the geographic coverage was now global and the scope extended to include nonmarine as well as marine taxa, however with the focus switched to genera in the first instance. The inclusion of the term "interim" in the project name was intended to convey the fact that in this instance, the focus would be on relatively rapid assembly of at least a "first pass" product for use by clients in the short term, which could then be improved as an iterative process over time, rather than waiting for every included data item to be as exhaustively checked as might be the case when the ultimate in data quality is required.

IRMNG GOALS

The goals of IRMNG as envisaged at the project outset were as follows:

1. Assemble as completely as possible a list of published genus names covering all domains of life, within a taxonomic hierarchy constructed as a set of logically consistent relationships (Allkin et al. 1992): e.g., a given genus cannot be in multiple families simultaneously; all taxa (apart from the root, i.e. "all life") must have a parent record which must be at a higher rank; all records are discoverable by traversing the taxonomic tree from top downwards; a parent cannot be deleted while it possesses "live" child records; etc. Non-current names (synonyms), when known to the compilers, should be pointed to their relevant valid name as information is at hand.

2. As many genus names as possible are then to be flagged marine or nonmarine (or both), extant or fossil (or both), for the use of clients such as OBIS and others.

3. As species data are readily available (for example as already compiled in the Catalogue of Life), add these to the system connected to relevant genera, and continue the flagging to specific level by either automatic propagation (where possible) or manual flagging as needed.

4. Genus names would be subject to reasonable scrutiny upon acquisition to remove duplicates, provide author citations in a consistent form, and insert names into the most appropriate position in the IRMNG taxonomic framework. For reasons of limited resourcing and to facilitate the assembly of an "interim" product within a realistic time frame, species names from generally reliable sources including the Catalogue of Life, regional compilations, and some museum databases would generally be accepted without additional checks.

5. Make IRMNG content available in bulk form as download files, either of the entire database content or of selected data items as appropriate, for upload into relevant client systems.

As the project progressed, additional aims were incorporated, including:

6. Make IRMNG content available for external query—and subsequently, web-based edit for designated content editors—via a publicly accessible website.

7. Include original publication information for names (at generic level in the first instance) as held in external resources: initially as "microcitations" (typically a journal or publication name, volume, and page in work, imported as a text string), later expanded with the capability to hold article titles and/or full (atomised) bibliographic citations in a separate "literature" module.

8. Include numeric identifiers for names as held in selected other systems (initially Aphia/WoRMS, subsequently ION: see below) to enable cross linkages to those systems via the web as desired. 2006 Catalogue of Life identifiers for species, originally uploaded to IRMNG, are not persistent and have subsequently been deprecated.

9. Include the ability to generate lists of homonymous names (i.e., where the same name has been used to denote different taxa on different occasions) via the web interface, in response to several user requests.

10. Include a "near" or "fuzzy" name matching process so that incoming misspelled names could be matched against correctly spelled target names, when held. This capability was found to be very valuable to users wishing to detect misspelled names on their own lists as well as providing the opportunity for the system administrator to detect and rationalise variant spellings of the same name already held within the system.

11. Include a range of online search methods to suit user requirements, such as search by full or partial scientific name, or by authority or year published, filter by higher taxonomic group, etc. Also, provide an option for remote users to input lists of names for bulk real-time matching to IRMNG holdings.

12. Provide the capability to report statistics on IRMNG data holdings at any time, including number of names held at different ranks, number of valid names vs. known synonyms vs. unresolved names in particular groups, numbers of marine/nonmarine and extant/fossil names per group, and more.

Estimations of the ultimate data volume are given later in this work, but at the outset it was considered sensible to plan for up to at least millions of names (i.e., comparable with the Catalogue of Life and other current systems), with the expectation that several hundred thousand of these would be genera.

SELECTED IRMNG DESIGN PRINCIPLES
What is a "taxonomic name" in IRMNG?

Since IRMNG aspires to hold only one record per "taxonomic name" (or more specifically, taxonomic name instance), it is necessary to articulate what is meant by this term in the IRMNG context. By taxonomic name instance we mean the scientific name used by a particular author at a point in a specific publication to formally describe a new taxon, thus representing a unique combination of taxonomic name, author and position of the relevant entry (typically page, or sometimes even line) in the particular cited work. It follows, therefore, that variations in the representation of an author name (such as presence or absence of initials, abbreviated form versus full form, etc.) and/or subgenus inserted into a species name do not comprise multiple taxonomic name instances under this definition, neither do variations of how the work itself is cited. As an example, in Patterson et al. (2010; their Figure 1) numerous possible variant representations of the name *Cyclotrachelus sodalis* are given including "*Cyclotrachelus sodalis* (Le Conte)"; "*Cyclotrachelus* (*E.*) sodalis (LeC.)"; "*C.* (*Evarthrus*) *sodalis* (LeC. 1848)" and many more; for IRMNG purposes these would all be rationalized to a single name instance, which in this case would be represented as *Cyclotrachelus sodalis* (Le Conte, 1848). Similarly, the names "*Acanthoperla* Cavalier-Smith" and "*Acanthoperla* Cavalier-Smith in Cavalier-Smith & Chao, 2012" would be considered the same name instance and represented by only one IRMNG record, as would *Ficus* Röding, 1798 versus *Ficus* Bolten, 1798, the two variants referring to the same work which has been ascribed variously to either Bolten or Röding in the past (the latter is now the accepted author of this work).

On the other hand, if the same (or different) author has published the same name as new on multiple occasions (perhaps first as a *nomen nudum*, followed by a subsequent valid publication) then this represents multiple instances of "name+author+position in cited work" and would be indexed accordingly as multiple records in IRMNG. This approach of creating only one record per "taxonomic name instance" contrasts with the concept of "name strings" as discussed, e.g., in Patterson et al. (2010), and as collected by indexers of such name strings including uBio[13], the Index to Organism Names (ION)[14], and the Global Names Index (GNI)[15], with the result that indexing operations such as

---

[13] http://www.ubio.org/
[14] http://www.organismnames.com/
[15] http://gni.globalnames.org/

those repositories potentially contain many more names-as-strings, each typically with its own designated identifier (such as a Life Science Identifier or LSID) than is the case for IRMNG.

The corollary of this is that incoming names to IRMNG may frequently need deduplication (multiple representations—of authorities in particular—being reconciled to a single preferred form), but that on occasion, care must also be taken not to amalgamate name instances which look similar (or even identical) but in fact refer to different taxa, i.e., homonyms in the broad sense. This includes both within-Code homonyms (homonyms *sensu stricto*) and duplicate names across different nomenclatural Codes ("transregnal homonyms" in Patterson et al., 2016), for example the same name used for both an animal and a plant, or a plant and a bacterium.

A practical example is provided by the genus name *Ceratium*, which has been used multiple times for different taxa, with each instance also potentially cited slightly differently in the different compilations that may be used as input to IRMNG. Rationalising and deduplicating these citations (name strings) is key to deciding how many IRMNG records should be created for this name. Combining data from five different sources (*Nomenclator Zoologicus*, *Index Nominum Genericorum*, *Index Fungorum*, Index to Organism Names and Global Names Index) we can arrive at the following rationalised IRMNG list (Table 1).

While rules mandate that homonyms within the same nomenclatural Code are not permitted (junior homonyms requiring a replacement name or *nomen novum*), no such restrictions apply to the same name being currently valid across different Codes and so (a single instance of) the same name can be concurrently valid in "plants" (historic usage, i.e. including algae and fungi), animals, and prokaryotes, (although probably not in viruses, whose genus names all end in "virus"). Thus in the *Ceratium* example given above, the oldest genus name (*Ceratium* Schrank, 1793, the dinoflagellate), if considered a zoological name, preoccupies subsequent identically named genera in zoology but not in botany, for which *Ceratium* J.B. Albertini & L.D. Schweinitz, 1805 remains a validly published name, with *Ceratium* Blume, 1825 an invalid (illegitimate) junior homonym.

Ranks in the IRMNG data structure

At the project outset and over the period represented by this report, for simplicity in data handling and also following the then-current edition of the Catalogue of Life, the IRMNG data structure above genus included only the "Linnaean" ranks, i.e, kingdom, phylum (= Division in botany), class, order and family, other intermediate ranks being dropped when supplied (but in some cases captured as an accompanying text remark). With the move of IRMNG content into the VLIZ data structure, the capability to easily add intermediate ranks exists and this has been taken up to a limited degree from 2017 on, commencing with the addition of subphyla such as Insecta, Crustacea and Myriapoda in Arthropoda, and Vertebrata, Urochordata and Cephalochordata in Chordata at this time. Protozoa and Chromista have also been revised, with intermediate ranks added thus far between the ranks of kingdom and class, following the scheme of Ruggiero et al. (2015). Over time the entry of further names at intermediate ranks can be expected; however, at time of writing the "Linnaean" ranks mentioned above remain the only ones consistently populated for all names. Also, for reasons of efficiency in data collection, ranks below the level of species have been ignored at the present time, although this may be revisited at a future date.

Treatment of subgenus names

At present, IRMNG does not include the rank of subgenus in its design concept, although this may be revisited in future versions. However, in zoology, names published as subgenera are deemed to be simultaneously available (published) at generic level—also at any other level in the "genus group"—via the principle of coordination (International Commission on Zoological Nomenclature 1999, article 43). As an example, the subgenus *Abyssopinna* created by Schultz & Huber (2013) within the genus *Pinna* Linnaeus, 1758 is also available (without change in cited authorship) if subsequent authors wish to use it at generic rank on account of a change in taxonomic opinion. Hence, within IRMNG, *Abyssopinna* is included as a published name at generic rank, although at that level it is listed as a synonym of its containing genus *Pinna*.

A second issue regarding subgenera is that, in zoology, at specific level it is legitimate to include a subgenus in parenthesis in between the genus name and the specific epithet, thus (taking the case of *Abyssopinna* as introduced above) the same species name can be represented as both *Pinna* (*Abyssopinna*) *epica* and *Pinna epica*. In

Table 1. IRMNG records for genus "*Ceratium*", with equivalents in selected other available data sources, namely *Nomenclator Zoologicus* ("A"), *Index Nominum Genericorum* ("B"), *Index Fungorum* ("C"), ION ("D") and Global Names Index ("E"). IRMNG numeric identifiers ("IRMNG ID") are allocated to the names on addition to the IRMNG system and are persistent (also are independent of identifiers for the same names in other systems). Note: where the IRMNG "preferred form" of the cited author and year differs from some or all of the sources used, the format has been selected and/or adjusted according to the principles set out in Appendix 2.

Taxonomic assignment	Verbatim records in sources A-E, arranged in one group per "taxonomic name instance" (refer caption for abbreviations used)	Equivalent name + authority as stored (in IRMNG-preferred form)	IRMNG ID	Remarks
class Dinophyceae (dinoflagellates)	*Ceratium* Schrank 1793 (A, D) *Ceratium* Schrank (B, D) *Ceratium* F. Schrank, 1793 (E) *Ceratium* Schrank, 1793 (E)	*Ceratium* Schrank, 1793	1274897	
class Myxogastrea (myxomycetes)	*Ceratium* Albertini & Schweinitz 1805 (A, D, E) *Ceratium* Albertini et Schweinitz (B) *Ceratium* Alb. & Schwein. (C, E) *Ceratium* Albertini & S. 1805 (D)	*Ceratium* J.B. Albertini & L.D. Schweinitz, 1805	1273955	Currently a synonym of *Ceratiomyxa* J. Schröter in Engler & Prantl, 1889
family Orchidaceae (angiosperms)	*Ceratium* Blume (B) *Ceratium* Blume, 1825 (E)	*Ceratium* Blume, 1825	1274128	Currently a synonym of *Eria* J. Lindley, 1825
family Pyralidae (moths)	*Ceratium* Thienemann 1828 (A, D, E)	*Ceratium* Thienemann, 1828	1274013	Currently a synonym of *Phycita* Curtis, 1828
family Pyralidae (moths)	*Ceratium* Gistl 1848 (A, E)	*Ceratium* Gistl, 1848	1274070	A later usage of *Ceratium* Thienemann, 1828 and thus also a synonym of *Phycita* Curtis, 1828
phylum Rotifera (rotifers)	*Ceratium* Agassiz 1846 (A, D, E)	*Ceratium* Agassiz, 1846	1274194	Currently a synonym of *Keratella* Bory St. Vincent, 1827

accordance with present IRMNG data conventions (Appendix 2), and again following the earlier practice of the Catalogue of Life, only the binomial version is used for species, and where incoming species names include the subgenus this will be removed before uploading the name to the system (it may however be checked to see whether or not the subgenus name is held, a least for zoological names, and where missing this should also be uploaded as a generic name in its own right).

By contrast, subgenera in botany and bacteriology are not automatically available for use as genera from their original publication; they are therefore not indexed as genera in IRMNG unless subsequently formally raised to that rank via a separate nomenclatural action. Botanical and bacteriological subgenus names are rarely encountered as a portion of a species name but where they are, the relevant element would be removed: for example "*Moraxella* (subg. *Branhamella*) *caviae*" would be entered simply as *Moraxella caviae*.

Unavailable names

In addition to available names (the zoological term, equivalent to validly published names in botany), IRMNG includes a component of unavailable names including some *nomina nuda*, original or subsequent literature misspellings, unjustified emendations, suppressed names, as well as bacterial names without standing in prokaryotic nomenclature (for definitions refer the relevant nomenclatural Codes), chiefly when these have been indexed in nomenclators such as *Nomenclator Zoologicus* or included in data compilations such as the Catalogue of Life. Such names are retained in IRMNG since they may appear in published literature and/or other taxonomic data compilations and may be required for correct assignment of accompanying information; whenever possible they are pointed to the current accepted name for the same taxon, when known. A few names in IRMNG are also classified as later usages: this applies when a name is indexed as new (for example in *Nomenclator Zoologicus*) but additional investigation indicates that the original valid publication of the name in fact occurs in an earlier work by the same, or a different author. As an example, in *Nomenclator Zoologicus* the name *Onychites* is credited to Zakrzewski, 1886 but, according to other published sources, in fact dates from a publication by Quenstedt, 1856. In this case, "*Onychites* Zakrzewski, 1886" is retained in IRMNG but cited as a later usage of *Onychites* Quenstedt, 1856, for which a separate entry is created.

Taxonomic uncertainty and divergent views

Taxonomy is not an exact science and the views of multiple authors regarding the taxonomic placement, rank, or status of a particular name or taxon do not always necessarily coincide, and may also change through time with advancing knowledge. Certain taxonomic information systems attempt to accommodate this by supporting multiple taxonomic views—an example being the present Encyclopedia of Life[16], which is capable of displaying classifications for any included taxon from multiple sources that may not always agree with each other.

By contrast, as is the case with other compendia such as Catalogue of Life and ITIS, IRMNG supports a single taxonomic view at this time which, if not always completely up-to-date, can be (and in some cases has already been) upgraded to follow arrangements generally recognised as "authoritative" for the group in question. Thus, angiosperm taxonomy, previously following the APG III treatment (Angiosperm Phylogeny Group 2009) has recently been upgraded to follow APG IV (Angiosperm Phylogeny Group 2016); the higher classification of fungi was updated in 2009 to follow the then-latest version of *Index Fungorum* (Kirk 2001-current), extant fishes follow the 2009 version of Eschmeyer's Catalog of Fishes (Eschmeyer 2000-current), and families of extant crustaceans, gastropod molluscs, and Coleoptera have been adjusted to follow recent treatments (De Grave et al. 2009; Bouchet and Rocroi 2005; Bouchard et al. 2011).

When alternative views exist and are known to the IRMNG compilers, these are typically captured in a taxonomic note appended to the record, for example the taxonomic remark for the Linnaean class "Aves" (IRMNG ID: 1142) presently reads: "Treated by some recent authorities, and Ruggiero et al., 2015, as a subclass of Reptilia; maintained as class at the present time in common with Catalogue of Life (2016), WoRMS (2017) and elsewhere". Changes introduced in the most recent treatments (for example combining or splitting families, genera or species) will not always be

---

[16] http://eol.org/

immediately reflected in IRMNG but, as resources permit, are hoped to be incorporated either directly from the literature or as they make their way into sources used as input to IRMNG, such as Catalogue of Life, The World Register of Marine Species, and others.

IRMNG attributes

As introduced above, a key initial driver for the compilation of IRMNG was the requirement for OBIS to distinguish between marine vs. nonmarine, and extant vs. fossil taxa. Accordingly, a system of "flags" was incorporated into the initial (2006-2014) version of IRMNG whereby a single "habitat flag" could be set to values corresponding to "marine," "nonmarine," "marine plus nonmarine," or null, i.e. not yet entered, together with a source for the relevant setting ("habitat flag source"). For extant/fossil status, a flag could be set to values corresponding to extant, fossil, extant plus fossil, or null. The concept of "marine" is intended to conform to the definition used in ERMS, namely "…broadly defined to include intertidal (littoral) and brackish water habitats, defined as up to the strandline or splash zone above the high tide mark and down to 0.5 ppt [parts per thousand] salinity in estuaries" (Costello 2000). This definition excludes salt marshes, which were considered in ERMS to be out-of-scope by virtue of being elsewhere included in terrestrial ecosystems. In IRMNG, "marine" is considered to exclude primarily terrestrial or freshwater species that may be found at sea in a non-obligate manner (for example certain ducks) but to include species which spend a regular portion of their life in the marine environment such as shorebirds and waders. Species which spend portions of their life cycle both at sea and on land (e.g. marine mammals, birds and reptiles that return to the land to breed) are assigned both marine and nonmarine status, as are species which alternate between marine and freshwater habitats at different stages of their life cycles.

From 2016 onwards, IRMNG supports additional options for habitat flagging in that "marine" is further divided into marine and brackish which can be separately assigned the states yes, no, or unknown, while "nonmarine" has been replaced by the categories freshwater and terrestrial which can be assigned equivalent states as required. To transfer legacy data most effectively from the initial version of IRMNG to the current one, names previously flagged "marine only" (which includes a component of

brackish water organisms) are represented as marine = yes, brackish = unknown, while former "nonmarine only" has been represented as marine = no, freshwater = unknown, terrestrial = unknown. While this representation is sub-optimal for some purposes, it still permits discrimination of marine vs. nonmarine taxa by interrogating the flag for "marine" = yes or no, and is capable of being upgraded through time as additional resources are available.

The IRMNG extant/fossil flag ("extant" being renamed "recent" from 2016 onwards) supports the options of recent only, recent plus fossil, fossil only, and unknown. "Recent" is defined as including taxa alive at any time since the beginning of modern scientific investigation, broadly interpreted as post 1500 A.D., thus including some species that have become extinct since that date (such as the dodo, *Raphus cucullatus*, last seen alive in 1662) but excluding species such as the extinct moas of New Zealand (of *Dinornis* and related genera), never seen alive by naturalists and known only from sub-fossil deposits older than c.1500 A.D.

IRMNG IMPLEMENTATION
Environment, programming language, virtual (web) and physical location

To address IRMNG goals, for the initial version a custom Oracle® database was constructed in August-September 2006 at CSIRO Marine Laboratories in Hobart, Australia, together with an application developed in the Oracle PL/SQL programming language to address the requirements for data query and web display, web editing, and other administrative functions. The database was made available for live query via the web in October 2006[17] and was subsequently upgraded to also incorporate input name parsing and fuzzy matching using the algorithm "Taxamatch" as described in Rees (2008a,b). Throughout 2006-2014 new content was added to the system (more detail given below) until 2014 when a process was commenced to transfer the database to the Flanders Marine Institute (VLIZ) in Belgium, as part of an agreed migration between the two institutions.

Over the next two years the content was progressively transferred to new data tables in Belgium and a new domain name was

---

[17] http://www.cmar.csiro.au/datacentre/irmng/

established for the project[18], with release of the new system at VLIZ announced in July 2016 after completion of suitable testing. Meanwhile, content acquired for the system post-July 2014 has been added to the copy at VLIZ but not to the CSIRO copy, which is thus effectively "frozen" as at that date and is planned to be deprecated once its functionality and use as a target for links from third party compilations is no longer required.

In its new location at VLIZ, IRMNG shares the same data table structure and code base as that developed for Aphia, the data store at VLIZ which supports WoRMS and other taxonomic databases (Vandepitte et al., 2015); however, the data content of IRMNG and Aphia currently remains separate. This means that (for example) IRMNG IDs are not interchangeable with WoRMS (=Aphia) IDs, IRMNG and Aphia displayed content for the same taxon may not be identical, and references are not shared between the two systems, although these aspects could change in the future.

The underlying database system at VLIZ is Microsoft SQL Server with a web interface implemented in the PHP programming language. The complete database design includes over 400 data fields spread over 81 tables (for a more detailed description see Vandepitte et al. 2015, also a diagrammatic summary of Aphia table relationships as shown via the WoRMS website[19]); only a subset of these are presently used for IRMNG, of which the principal ones are described in the next section. The actual migration process for the IRMNG data was non-trivial, and involved comparison of the existing IRMNG and Aphia data structures, mapping of the relevant data fields, import of the data, and then an appraisal and iterative adjustment process to ensure that fields had been mapped in the most meaningful way and no content deemed important was lost during the migration. Separately from the actual data migration, new web interfaces for general information, data search, and online editing of both taxonomic names and relevant literature were constructed (based on pre-existing templates but customized to suit IRMNG requirements), and tested / adjusted as needed prior to public release.

With the move of the database to VLIZ, a number of additional search options (some via "advanced search") developed for Aphia/ WoRMS but not previously available in IRMNG have also been implemented. New features include, among others, a browsable taxon tree; search by IRMNG ID as well as taxonomic name; search by taxon status (i.e., accepted, unaccepted, *nomen nudum*, etc.); search by habitat and/or extant/fossil flag status; limiting a search to above or below a designated taxon rank; search by note type and any particular included text desired; and literature (source) search, in addition to filter by higher taxon and search by scientific name and/or name author, which were previously also offered in the original search interface at CSIRO.

By default, searches via the new interface execute across, and return results from, multiple ranks simultaneously, an improvement on the original capability at CSIRO which only returned results from the same taxonomic rank as specified by the user. A previous capability for input of multiple names to the standard "search" box has been replaced by a more flexible Taxon Match tool, based on that previously developed for WoRMS, via which a user can submit up to 1500 rows of data (taxon names) for matching and be returned a file in CSV format containing a flexible, user-selected range of output variables including any or all of IRMNG_ID, scientific name, authority, accepted name, higher classification, quality status, taxon status, environment, and citation (= publication details for the name). Additional description of the detailed operation of this tool is available via the Taxon Match Tool User Manual[20]. Examples of the current (May 2017) IRMNG "basic search" interface, with a sample search result and example taxon page, are given in Figures 1-3, with the additional options offered via "advanced search" in Figure 4.

IRMNG data fields

IRMNG holds a set of data fields associated with every taxonomic name instance, plus (in additional tables) details for references/sources, values displayed in picklists used at data entry time, etc.; in addition, it holds internal administrative data on persons who have made or can make changes. Available data fields associated with each taxonomic name instance are shown in Table 2.

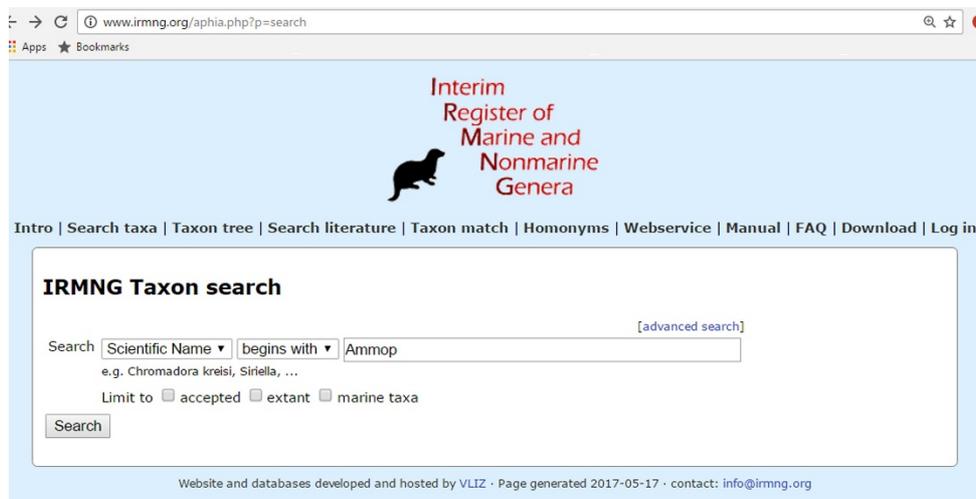Figure 1. IRMNG "basic search" interface at VLIZ, current (May 2017) version (http://www.irmng.org/aphia.php?p=search).

Figure 2. Initial portion of the search results page produced by the search as specified in Figure 1. The search operates simultaneously across all ranks (unless otherwise requested), includes names from botany, zoology bacteriology and viruses as appropriate, and also both extant and fossil taxa, the latter indicated with the dagger symbol (†) as a suffix.

Figure 3. Example generic level taxon page produced by clicking on the link to "*Ammopemphix* Loeblich, 1952" from the search results shown in Figure 2 (a small number of cited sources are omitted for clarity). Note that, as is the case with the majority of genus-level pages in IRMNG, the information displayed is aggregated from multiple sources, in this case *Systema Naturae* 2000 and for the initial genus name and authority, *Nomenclator Zoologicus* for the original publication details, genus spelling and authority verification, and a taxonomic remark, WoRMS for the family allocation, status as an accepted (current) name, and WoRMS identifier (included in the pre-formatted link as displayed), Ruggiero et al. (2015) for the classification between kingdom and order, and the ION database for the ION identifier which is used to create relevant deep links to both ION and the BioNames database. The two displayed child species are sourced from Aphia2006/ERMS and WoRMS (2013 version), respectively. Additional information on sources used is given in Appendix 1.

Figure 4. IRMNG "advanced search" interface at VLIZ, current (2017) version. Name status "accepted" corresponds to available names (in botany: validly published names) not presently known to be synonyms, in other words both valid (current) names plus a subsidiary component of names not yet scrutinized for taxonomic status.
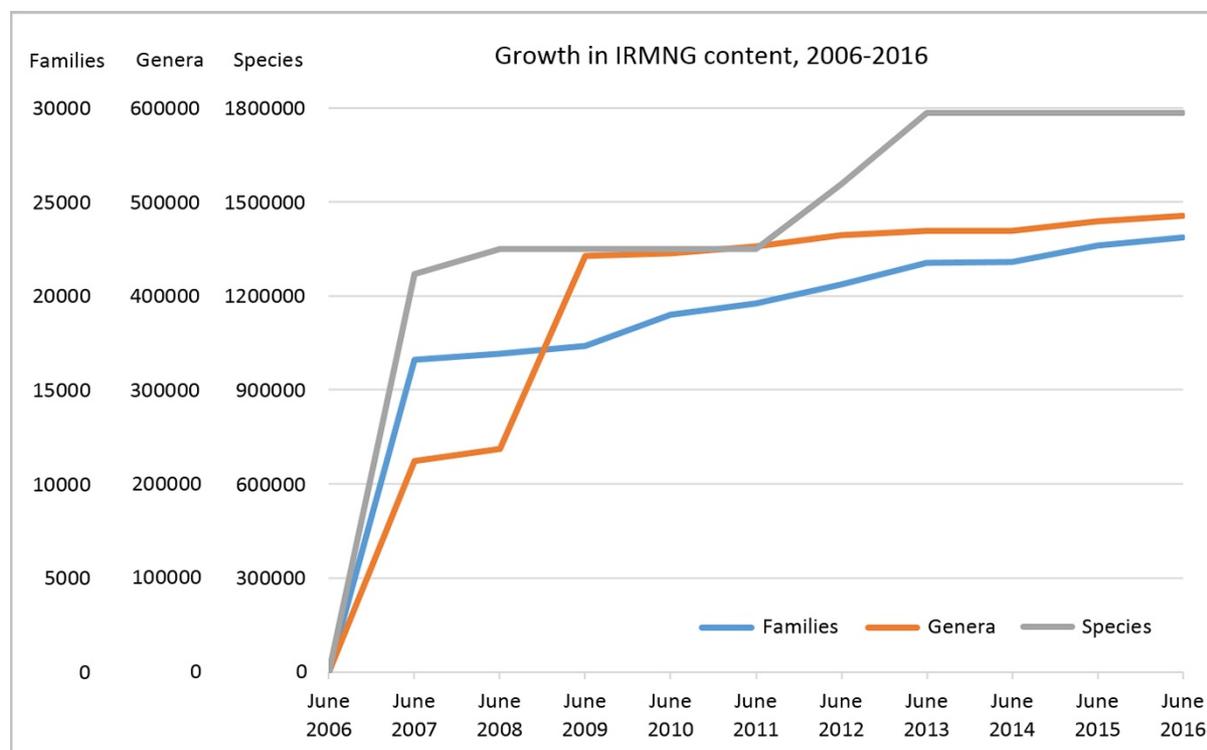

Figure 5. Growth in IRMNG content, 2006-2016, as numbers of records held at family, genus and species levels.

Table 2. IRMNG data fields for taxonomic names. "Required" fields are those that must be populated before a new IRMNG record cannot be saved; "desirable" are fields that should be completed when information is available, either at initial upload or as a subsequent activity.

Priority	Field name/s	Remarks
Auto-created record attributes	IRMNG ID	Unique, numeric identifier (auto-allocated, as next available number).
	Record creator and creation date	
	Record status	By default, "active"; can subsequently be altered to either "in quarantine" or "deleted" if required.
Required fields	Scientific name	If genus rank and above, the uninomial name; if species rank, the specific epithet only (to be combined with its associated genus for record display, searching, and data export).
	Rank	Taxonomic rank of the name, selected from a set of pre-defined options (kingdom through species)
	Parent	Parent taxon, as taxonomic name selected from a picklist (then held internally as the relevant IRMNG ID), permitting placement in the taxonomic hierarchy. Top-level taxa, presently kingdoms, have the single parent "Biota", which has IRMNG ID =1.
	Status	Name status i.e. one of the following options as per the present WoRMS standard: • accepted • unaccepted • *nomen nudum* • alternate representation • *nomen dubium* • temporary name • *taxon inquirendum* • interim unpublished.
	Habitat flag/s, i.e., marine, brackish, freshwater, terrestrial	For names entered 2016 onwards, at least one of these four flags must be set to either "yes" or "no" (permitted values for each flag are "yes", "no" or unknown").
	Accepted name	If the taxon status is entered as "accepted", this is by default the taxon name entered; if the name status is "unaccepted", the currently accepted name is required, selected via a picklist drawn from names already in the system.
Desirable fields	Authority	The authorship associated with the scientific name, in IRMNG-preferred form (refer Appendix 2)
	Extant/fossil status	Permitted options are recent only; fossil only; recent + fossil; unknown.
	Linked source	A repeatable link to an IRMNG reference (entry in "sources" table via relevant Source ID, see below),

categorized under one of the following headings:

- original description
- basis of record
- additional source
- source of synonymy
- redescription
- new combination reference
- status source
- toxicology source
- taxonomy source
- ecology source
- identification resource
- subsequent type designation
- misapplication
- original description (unavailable nomenclaturally)
- emendation (re-diagnosis of genus)
- verified source for family
- verified source for genus
- current name source
- extant flag source
- habitat flag source
- context source

A number of these fields had no equivalent in the original (CSIRO) version of IRMNG and are not populated at the present time, but are available for use from 2016 onwards. "Original description" corresponds to the "publication" field in the initial version of IRMNG, and may be populated with a full bibliographic citation or simply with an abbreviated "microcitation" (e.g., journal, volume, page) as shown in the example given in Figure 3 above. "Basis of record" is used as required to identify the source from which the name was acquired, as per the examples given in Appendix 1.

| | Unaccept reason | A free text field to supply more information to the user on the reason for a given name having the status "unaccepted"; examples: junior synonym, junior homonym, name published in suppressed work, misspelling, etc. |
| Supporting information as available | Notes | Additional notes (e.g., nomenclatural or taxonomic comments, or information on habitat or geologic range), either as present in the source used, or added by an IRMNG editor. Notes take the form of a repeatable free text field (optionally with an associated source), assigned to one of the following categories: |

- Classification
- Descriptive information

- Habitat
- Nomenclatural status
- Nomenclature
- Subfamily
- Taxonomic remark
- Type species as cited
- Validity

Nomenclatural status is completed where applicable, for example: *orthographia* (misspelling), *nomen novum*, *nomen nudum*, etc. "Nomenclature" defines the nomenclatural Code governing the use of the name, chosen from ICZN, ICBN (historic name for the current botanical Code or ICNafp), Bacteriological Code (BC) and Virus Code (Vir). "Type species as cited" is provided to hold any text providing this information (e.g. as available in an IRMNG data source) without the requirement for the name to presently exist within the IRMNG system, and could later be converted to a relevant IRMNG ID as available. "Taxonomic remarks" may include any comments regarding alternative taxonomic views of the taxon in question, remarks on the availability of a replacement name if the name is a homonym, presentation including cited authorship, etc.

ID in other system/s	System-specific identifiers for the same name in other systems, currently comprising a dedicated field for ITIS Taxonomic Serial Number (TSN), plus a repeatable "Links" field presently holding web links (including relevant identifiers) to the World Register of Marine Species (via Aphia ID), and the Index to Organism Names, and BioNames, both of which can be addressed using ION IDs.
Type species	For genera, indication of the type species (selected from a picklist of relevant names already entered), plus optional type designation method selected from a picklist.
Original name	For species (e.g., new combinations), indication of the original name on which the new combination is based.
Gender	For genera and species, indication of the grammatical gender of the relevant genus name or epithet, selected from the options masculine/feminine/neuter/ unknown.
Updater name, date, aspect changed	Record update history, when changed since original entry (field repeatable as needed). Details on specific aspects changed are held in the database and can be reviewed by administrators as required.

It should also be noted that the order of record creation proceeds downwards through the taxonomic hierarchy; for example, if a new family, genus and species are required, the family must be created first in order that it can be cited as the parent of the genus, and the same for the genus and species records. Further, since every record "knows" the identity of its parent in the IRMNG system, lists of child taxa at any level are generated automatically for display as part of a taxon page (e.g., species of a particular genus) without any requirement for data entry at the level of the parent.

"Active" records are those displayed in response to user searches. The status of a record can be changed (by a relevant IRMNG editor) to either "in quarantine" or "deleted" as required. Records in quarantine are basically in-progress, "work" records requiring more effort or scrutiny before a decision is made to make them publicly viewable, while records may be deleted once a decision has been made that they are no longer required, for example being duplicate entries, "temporary" parent records that no longer have any children, or records based on erroneous external data that are not required to be kept. Deleted records remain on the system so that first, their numbers are never re-used, which could lead to confusion; second, their current status and reason for deletion can be tracked for administrative purposes, and incoming queries against them can be redirected to a designated active record; and third, they can be reactivated (undeleted) if required, for example if the record was deleted in error, or if the reason for its deletion ceases to apply.

IRMNG sources are held in a dedicated module labelled "literature" which is available for user search via its own dedicated interface linked from the IRMNG home page. An individual source can be either a free text statement such as "Cavalier-Smith, 1992" (typically a legacy entry from the previous version of IRMNG), a microcitation (as per the example shown in Figure 3), or a fully atomised bibliographic reference, and can optionally include an online link to a more complete version of the cited work on an external accessible site. IRMNG sources have a numeric ID which is allocated on source creation and may be linked to multiple statements about multiple taxa: for example the source "Nomenclator Zoologicus" is presently linked to >300,000 such statements.

With the move of IRMNG to VLIZ and the data structure developed for Aphia, certain other fields have become available for use within the IRMNG data structure, namely vernacular names (and relevant language), distribution, specimen details for a species, feeding type, images, and "contexts", the last being a set of "tags" controlling within which of the multiple systems supported by Aphia the relevant record is to be displayed. Currently, none of these fields is populated for IRMNG but they may be utilized in the future.

Since names may be uploaded with only the "minimal" information in the first instance, an important aspect of ongoing IRMNG population (in addition to adding new names not previously held) is to add or upgrade missing information items for names already within the system. Accordingly, a web-based data edit interface is provided (to suitably authorized editors) which supports both the creation of new records, and upgrade or alteration of any aspect of existing records. Data can also be added or amended as bulk operations on request to the database administrators, which can be preferable where particular tasks would be inefficient for editors to perform via the online edit interface on a record-by-record basis; examples of the latter might be supplying a list of names/IRMNG IDs with new information to be entered against each (such as a habitat or extant/fossil flag where these are currently missing), or requiring the move of a substantial number of child taxa from one parent to another.

Population of the IRMNG system

The initial task in developing IRMNG in 2006 was to construct a higher taxonomic framework (kingdom through family) into which incoming lists of genera and/or species could be placed. Concentrating on the "Linnaean ranks", the family-level treatment of Parker (1982) was used for extant taxa, as the most comprehensive then available despite being some decades out-of-date (a few minor groups accidentally omitted from the latter were added as required from other sources). Subsequently, fossil-only families were also added using a digital summary of data from Benton (1993). Input of the family treatment of Parker was facilitated by digitization of the relevant portion of the printed work in collaboration with the team at the SeaLifeBase project[21], while for the fossil families, a spreadsheet version of the holdings from Benton

---

[21] http://www.sealifebase.org/

(1993) was used as cited in Appendix 1. Habitat flags were created for all families based on indications in both works, those in the Benton compilation already being present within the electronic file used.

With the initial family and higher taxonomic framework in place (subject to later revision as required), available sets of genus and/or species names were formatted for addition to IRMNG and progressively uploaded to the system (refer Appendix 1 for full details). In addition to the major sources listed in Appendix 1, a large number of smaller compilations and individual papers have been consulted, either as sources of additional names or to obtain supplementary information about specific name instances. As sources are consulted, their details are added to the literature module as described above, from which their details can then be displayed on relevant taxon pages.

Over the ten-year period covered by this report, IRMNG holdings have grown to 486,652 genus names, of which 362,597 are presently designated "accepted", within a higher taxonomic structure consisting of 23,243 families, 3,107 orders, 587 classes, 162 phyla and 7 kingdoms, including a small number of higher taxon names currently regarded as synonyms. The period of most rapid growth for families was 2006-2007 including uploads from Parker (1982) and Benton (1993), for genera 2006-2009 (including initial uploads from *Systema Naturae* 2000, Catalogue of Life—the latter without authorities, *Index Nominum Genericorum* and *Nomenclator Zoologicus*), and for species 2006-2007 (corresponding to the addition of over 1.2 million names from the Catalogue of Life), supplemented by a boost between 2011 and 2013 corresponding to the addition of more species names from the Hallan Biology Catalog and the 2013 edition of WoRMS. Overall growth in IRMNG holdings over time at family, genus and specific levels is illustrated in Figure 5.

Figure 6 shows a breakdown of the genus-level content as at December 2016 by major taxonomic group, also indicating the proportions of genus names within each group flagged extant (= recent) only, extant + fossil, fossil only, and unknown extant/fossil status as held at that time.

IRMNG DATA LOADING, STANDARDIZATION, AND QUALITY CHECKS

For the initial components of IRMNG content to be uploaded, the data were formatted to suit the IRMNG data structure and then added to the system, with a unique IRMNG identifier created for each name. As information was available, synonyms and some unavailable names were pointed to their equivalent accepted name (valid taxon) as appropriate. Habitat and extant/fossil flags were added whenever possible by flagging entire families using available information from the compilations by Parker, Benton, or elsewhere, and otherwise as discrete exercises based on a range of published literature and/or web sources.

Subsequent batches of candidate data for addition to the system are first checked to see if names present are already held on the system before the remainder are prepared for upload by adding the minimum required fields, plus as many of the desirable/optional fields listed in Table 2 as can easily be populated. Some names in incoming lists might not be loaded if, for example, they are misspellings or simply variants of names already held, while on occasion, names apparently already held will still be loaded if they represent a new instance, for example a previously unknown homonym of a name already in the system, possibly even at a different rank (such as the protist phylum *Sagenista* Cavalier-Smith, 1995, homonym of the genus *Sagenista* Bohart, 1967 in Hymenoptera).

The matching of candidate names for upload with names already held on the system may be complicated by varying degrees of authority match, for which a decision must be made on a case by case basis aided on occasion by prior experience and/or supplementary research. For instance, on some occasions apparently dissimilar cited authorities can in fact represent the same publication instance, for example *Ficus* Röding, 1798 vs. *Ficus* Bolten, 1798 (in this case, both Röding and Bolten have been cited as the author of the work in question, so only one taxonomic name instance is involved). Conversely, apparently similar authorities may represent different name publication instances, for example two instances of *Sosxetra* Walker, 1862, but with slightly different publication details (in Lepidoptera and Hymenoptera, respectively), or *Enhydrus* Macleay, 1825 (a beetle) vs. *Enhydrus* MacLeay, 1925 (a mammal).

To assist in authority comparisons, where names are identical, the "authority match" component of Taxamatch (Rees 2014) can be employed to quantify and pre-sort names according to the degree of authority similarity, as
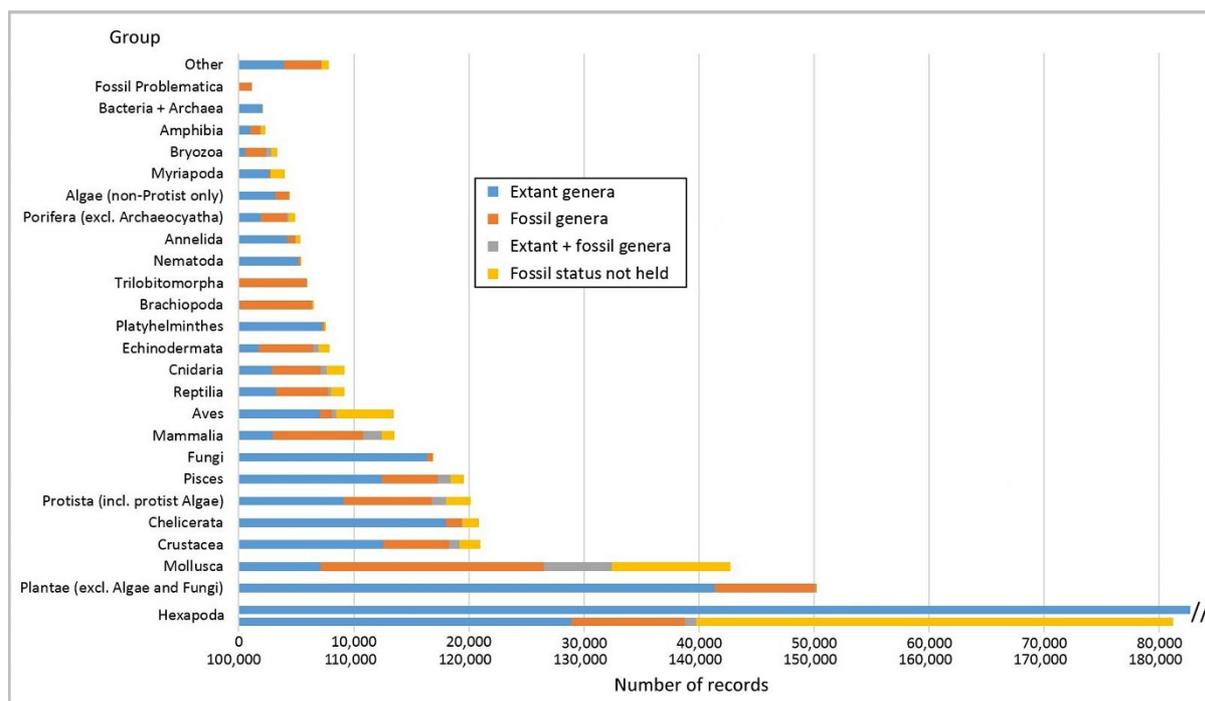
Figure 6. IRMNG genus holdings at December 2016 by major taxonomic category, with breakdown by extant/fossil status as held in the system. Extended scale at bottom of chart (100,000-180,000) applies to the second bar of Hexapoda only. Categories displayed represent a mix of formal plus informal groups as devised for the CSIRO implementation; to replicate these via the new search interface may require some component groups to be summed (for example to produce totals for non-protist Algae, "Pisces", and Plantae as displayed).
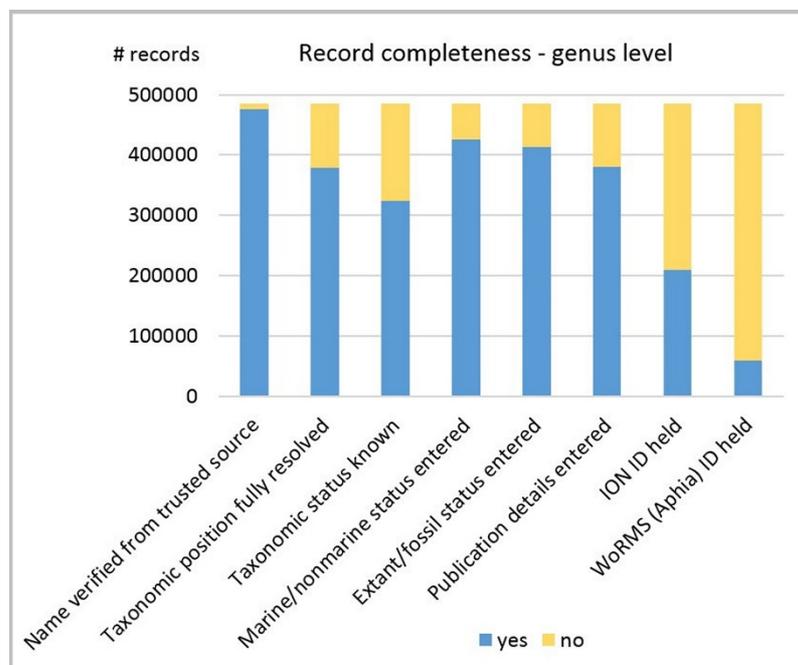

Figure 7. Breakdown of IRMNG genus holdings at December 2016 by key attributes entered. For additional detail refer text.

a precursor to manual scrutiny of the differences thus detected. In practice, higher calculated degrees of author similarity (for example >0.7 on a 0-1 scale) suggest that two name instances may be the same (enabling generally more rapid scrutiny), with the converse applying for lower levels of similarity (e.g., <0.3), suggesting that the name instances may well be different, again assisting the speed of a potential accept/reject decision. Names with an intermediate degree of author match are set aside for more detailed review as required.

Incoming data to IRMNG are typically reviewed for errors and/or inconsistencies, and may be edited as needed to suit in-house conventions or correct obvious data errors as required. Present conventions used for data standardization in IRMNG are listed in Appendix 2. Where possible, the bulk of the desired editing, such as restoring missing diacritical marks on author names, expansion of abbreviated botanical authors and addition of missing dates, is carried out prior to data upload. Another area of quality assurance is provided by the IRMNG Taxon match tool (which again uses Taxamatch) which, when presented with an incoming name not found in other reputable sources, can return suggestions as to names already in the system of which it may be a misspelling.

Once loaded, the availability of more extensive sets of comparative IRMNG data provides opportunity for additional scrutiny via review of names sorted binned in different ways (such as alphabetic sort by scientific name or author, review of all genera in a family or species in a genus), which can further assist in revealing inconsistencies or errors (such as exact or near duplicate entries originating from multiple sources, the same author name cited in multiple ways, or an inconsistency in cited dates) which can then be addressed as resources are available.

IRMNG COMPLETENESS AND LIMITATIONS
Numbers of records held, versus anticipated "complete" data holdings

In an ideal situation it would be helpful to know in advance how many published family, genus, and species names exist so that first, the eventual size of the system when fully populated could be gauged, and second, the degree of IRMNG completeness could be estimated at any given time. Parker (1982) gives a list of approximately 6,900 extant families, with a further 3,600 fossil-only families listed in Benton

(1993), both listings considered essentially complete by the relevant authors at those times, and all uploaded to IRMNG as described above. A further 12,700 family names (including 1,400 presently regarded as synonyms, plus an unknown number of variant spellings or synonyms not yet detected) have subsequently been added from a range of sources, suggesting that IRMNG is likely to be quite complete for accepted names at family level. A comparison of IRMNG extant, accepted family names with the 9,650 extant families listed in Ruggiero (2014) would also be instructive but has yet to be undertaken. It should, of course, be noted that a large number of family names proposed in the past are presently regarded as synonyms, but since IRMNG does not aspire to include these exhaustively, the extent of this situation is not relevant here.

For genera, the most complete vetted (i.e., deduplicated) recent listings are the approximately 357,000 zoological names indexed by *Nomenclator Zoologicus* up to 2004, together with the 69,000 botanical genera in *Index Nominum Genericorum* 2012 version (minus overlap with *Nomenclator Zoologicus* in 2,500 cases), plus prokaryote genus names from the Catalogue of Life and the List of Prokaryotic names with Standing in Nomenclature (2,200 names in 2008) and virus names from the Catalogue of Life and the ICTV Virus database (420 names in 2011). These totals include both accepted and unaccepted names (valid names plus synonyms) and also, in the case of *Nomenclator Zoologicus*, a component of unavailable names including published misspellings and *nomina nuda*. Together, these total approximately 426,100 names, to which should be added first, totals for names within scope for, but missed by, the nomenclators as given above, and second, the number of names published since the cut-off dates for those compilations and their respective versions. Estimates for the first of these are not available; however, IRMNG thus far contains approximately 22,000 dated names in zoology and botany published before 2004 plus a further 1,000 undated names, not in relevant nomenclators. If it is reasonable to suggest that over the past ten years, IRMNG has uncovered perhaps 50-70% of such missing names, then a further 10,000-20,000 may exist to be found and indexed through time.

Over the period since 2004, an estimate of IRMNG completeness is somewhat easier to

determine since annual publication rates of new genus names have been relatively stable at around 2,500 per year for all groups (IRMNG data, 2000-2009 average) so, for the period 2010-2016 inclusive a total of around 17,500 published names would be expected. IRMNG presently holds 9,831 published names for the period 2010-2016 leaving an estimated shortfall of around 7,700 published names to be acquired in this respect. Overall, the shortfalls estimated comprise around 18,000-28,000 generic names suggesting that a total of around 510,000 published genus names may exist, with present IRMNG holdings (487,000 genera) therefore comprising an estimated 95% of all such names published.

At specific level, estimates of accepted (valid) names do exist, namely those of Chapman (2009) for extant taxa (1.9 million) and Raup (1986), Alroy (2002) and Prothero (2013) for fossils (250,000-300,000 depending on source consulted), giving a present published total of up to 2.2 million accepted species as at 2009, also increasing by up to around 20,000 new names per year[22]. In the context of "all names" this number must be expanded to include both synonyms and previous or alternative binomials (i.e., genus+species combinations) now outdated by subsequent genus transfers. The precise extent of the "synonym problem" is unknown. In perhaps the most extensive, expert-vetted compilation to date for a major portion of the living world (vascular plants and bryophytes), version 1.1 of "The Plant List" [23] contains 350,699 species names listed as "accepted", 470,624 as known synonyms, and a further 242,712 as "unresolved", the majority of which are probably also synonyms. Such values may be typical across other portions of the taxonomic realm, for example Benton (2008) reports a synonymy/error rate within named dinosaur species of 48.2% (not including alternative combinations) while in sources considered by Patterson et al. (2016) a synonymy rate of around 3 synonyms per accepted name is regarded as the most typical value. Using this estimate, in addition to the approximate figure of 2.2 million species a further 6.6 million synonyms may exist, giving a total of approaching 9 million species names published to date. On this basis at 1.78 million species names IRMNG presently

holds around 20% of estimated published species names at this time, leaving a further approximately 7 million to be collected should this be an eventual design goal.

Overall record completeness (at generic level)

Completeness of IRMNG data holdings should not just be assessed by simple numbers of records held, but also by the relative completeness of key attributes of those records. For this purpose, relevant data are presented in Figure 7, with remarks against the individual measures used given below.

Regarding the attributes shown in Figure 7, the following notes are applicable:

- *Name verified from trusted source:* this applies to incoming names which can also be found in major nomenclators for the group in question, or have been verified from their citation in the primary literature, either from the original description or as part of a formal taxonomic treatment. A small number of genus names (approx. 10,000) has entered IRMNG from sources not considered "trusted" in this sense, such as museum databases or third-party compilations, and do not match entries in major nomenclators on a first pass; such names are identified as candidates for further investigation and will either be accepted or rejected for continued IRMNG use in due course.

- *Taxonomic position fully resolved:* this means that a genus has been allocated to a known family; names not yet allocated to a family are associated in IRMNG with the next higher category for which a placement is available, such as "Mammalia (awaiting allocation)", "Arthropoda (awaiting allocation)", etc. This attribute is less than 100% complete because one significant source in particular (*Nomenclator Zoologicus*) does not contain family allocation for its included names, meaning that that the latter must be backfilled from other sources.

- *Taxonomic status known:* this covers names that are determined to be either accepted names for current taxa (valid name in zoology, current name in botanical usage), or unaccepted names (taxonomic synonyms, or unavailable names for other reasons as determined in relevant nomenclatural Codes). Names acquired solely from *Nomenclator Zoologicus* lack this information, which again must be backfilled from other sources.

- *Marine/nonmarine status entered:* once again, this information is not available via standard nomenclators; however, where the taxonomic

---

[22] http://www.esf.edu/species/
[23] http://www.theplantlist.org/

position has been resolved to family (or higher taxon in some cases) this can frequently be allocated by inheritance of the relevant status from containing higher taxa.

- *Extant/fossil status entered:* major nomenclators such as *Index Nominum Genericorum* and *Nomenclator Zoologicus* do contain indicators of extant/fossil status; however, those in the latter compilation are not exhaustive so that while a fossil indication can be relied on, names without a fossil flag have been found to be not exclusively extant. Therefore, this attribute is not entered in IRMNG as "extant" until names sourced from *Nomenclator Zoologicus* lacking a fossil flag therein have been independently checked in other sources, leading to a lesser degree of completeness for this attribute than would otherwise be the case.

- *Publication details entered:* animal and protistan names from *Nomenclator Zoologicus* do have associated publication details (as microcitations), which is the chief function of such a nomenclator and these have been carried through to the relevant IRMNG records. Equivalent publication details for botanical names were not included in the original download file from *Index Nominum Genericorum* provided for IRMNG use in 2007 and so were not uploaded at that time; however, they are available via the latter's website and may be added to IRMNG as an enhancement at a future time.

- *ION ID held:* ION covers the zoological subset of taxonomic names and creates ION identifiers for any names encountered during the creation of the Zoological Record compilation which indexes both the primary literature (in this regard, descriptions of new taxa from 1864 onwards) and the secondary literature. Consequently it has created a large number of identifiers (over 3.5 million at the present time, with perhaps one tenth of these being genera); however, many of these are essentially duplicates in the IRMNG sense (different IDs for versions of the same name differing only by slight variants in their cited authority), or in some cases represent literature misspellings. IRMNG has harvested only the ION IDs which have been created from the original published descriptions of the taxa involved since these are of the highest quality and also of the most benefit to users, in that following the ION ID as a deep link to either the ION or BioNames database[24] will lead to a more complete citation of the original description of the taxon involved. At

present, ION IDs are held for 43% of IRMNG genera, but this increases to 66% if those not in scope for ION (non-animal names, and animal names published prior to 1864) are excluded.

- *WoRMS (Aphia) ID held:* Aphia IDs were uploaded to IRMNG for linking purposes as part of a cross mapping exercise in 2013 and this process is intended to be repeated at intervals as content is added to both systems. Since the WoRMS system is currently limited to marine taxa in the main, also with relatively few fossils, the proportion of IRMNG genera with this attribute populated will not be likely to exceed about 12% (the proportion of accepted, extant eukaryotic species currently estimated to be marine using data from Appeltans et al., 2012 and Chapman, 2009), presuming that genus representation is similar to that for species and that trends for synonyms follow those for accepted names. In practice the proportion will be less again to the degree that fossil names are currently under-represented in WoRMS. At present, WoRMS/Aphia IDs are held for 12% of IRMNG genera, but this increases to 48% if those not in scope for WoRMS (IRMNG genera not flagged marine = yes) are excluded.

Data are not presented above for species because this rank is not a primary goal for IRMNG at the present time, but since the majority of species records currently held are associated with genera already verified, taxonomically resolved, and possessing relevant habitat and extant/fossil flags, it can be expected that overall, species records are more "complete" than the values shown for genera for the displayed attributes with the exception of publication details and ION IDs. Publication details for species were generally not available in the sources used to compile IRMNG to date, though some will be available via deep links to ION and/or WoRMS, while ION IDs have only been entered where these link to original publication details for the name in question, thus excluding new combinations which (in zoology) are generally not tracked via changes in authorship and associated bibliographic citations.

Other known limitations and data gaps

Apart from the indications of present database completeness presented above, some other present limitations of IRMNG content should be noted, in particular:

---

[24] http://www/bionames/org/

- Despite reasonable efforts to avoid this, some genus records may contain inaccuracies in assigned habitat and extant/fossil flags, and in the currency or correctness of stated synonymy assertions and family assignments.

- Records between phylum and generic level (i.e., class, order, family) have not been subject to the same level of scrutiny as genera and may contain errors (e.g., misspellings and out-of-date names) or other inconsistencies.

- The higher taxonomy of many groups may be outdated to varying degrees, and has only been revised to follow the latest published sources for selected groups (such as extant angiosperms) at this time.

- Updates to the status of IRMNG species can lag behind those for genera, so some "unaccepted" genera may still contain species flagged as "accepted" at this time. As an example, over 100 "accepted" species of the genus *Michelia* Linnaeus, 1753 were uploaded from the 2006 version of Catalogue of Life; however, according to more recent sources, that genus is now treated as a synonym of *Magnolia*. Transfer of the affected species has not yet been made in IRMNG, and in any case would not be done until a relevant source is located in which the relevant revised combinations are supplied.

- To the extent that a particular group contains presently unallocated genera (for example, phylum Mollusca currently contains 8,600 such names, order Coleoptera 3,800, and class Reptilia 1,900), listings of genera for its families may be incomplete, with some otherwise valid family names having few or even no listed genera in the IRMNG data structure. (Over time, this situation should improve as "unallocated" IRMNG genera are scrutinized and further taxonomically resolved).

- Species records have received less scrutiny than genera at this time and are known to include some duplicates and misspellings as uploaded from the various data sources utilized. An effort will be made to further deduplicate and rationalize this element of IRMNG data holdings in the future.

Data gaps in IRMNG at generic level principally comprise a proportion of genera published since 2010 (2009 for prokaryotes and fungi), with coverage for animal genera continuing through to 2014 at a decreasing level. At levels above genus, coverage should be essentially complete with the exception of a small number of families or other higher taxa recently established or resurrected. As previously noted, at specific level IRMNG does not presently aspire to completeness and in general, gaps exist in the terrestrial area for some groups not covered in the IRMNG sources used to date (marine species are generally covered via recent updates from WoRMS) and, more particularly, for fossils in many groups, although the latter could be addressed in part via future imports of relevant data, for example, as contained in the PaleoBiology Database and elsewhere.

HOMONYMS IN IRMNG

Homonyms in the strict sense are recognised only within a particular nomenclatural Code, for example "botanical" groups (McNeill et al. 2012), zoological names (International Commission on Zoological Nomenclature 1999), and prokaryotes (Lapage et al. 1992; Parker et al. 2015) with the same name being legitimately available (and therefore not technically a homonym) between Codes. For example, the genus *Ficus* is the current name for both a flowering plant and a gastropod, while *Peranema* is both a protist and a fern. Homonymy in the strict sense also excludes unavailable names such as *nomina nuda* and subsequent misspellings, which may nevertheless be found in the literature used as identifiers for taxa. In IRMNG, therefore, we use the term homonym in an expanded sense to cover any multiple instances of the use of the same name for different taxa whether within or between Codes (also whether or not available), so that users can be alerted to sources of potential confusion and given a pointer to the fact that the taxonomic placement of a particular named taxon in external data may require to be checked further.

A capability has accordingly been created within IRMNG to generate lists of homonymous (i.e., duplicate) names at any rank which is constructed on-the-fly from the database, in other words, as a duplicate name is entered to the system (or removed) the list is automatically updated without any requirement for manual curation. Present statistics indicate that there are around 77,000 homonymous genus name instances (including small numbers of misspellings and *nomina nuda* in addition to validly published names), plus around 190 homonymous family names as encountered with incoming data during IRMNG construction (the list at family level is by no means exhaustive since many older / non-current family names are not presently included in IRMNG). A set of homonymous species names can also be

generated on demand; this list is quite extensive (around 100,000 records / 30,000+ names, which may however also include some entries requiring deduplication) but the number reduces to just 160 presently known to IRMNG if it is restricted to species-level homonyms associated with different genus instances (e.g., *Abronia aurita* the reptile, versus *Abronia aurita* the angiosperm).

An example from the present list of known family-level homonyms is shown in Figure 8; lists for family, genus and specific level can be generated on demand via the "homonyms" link indicated on all present IRMNG pages, or directly via this URL[25].

The incidence of genus-level homonymy can be illustrated by inspection of a list of IRMNG genera names simply sorted alphabetically (example shown in Figure 9), in which the occurrence of names sharing the same spelling is readily apparent. Furthermore, such sets are not limited to name pairs: the case of *Ceratium*, with 6 instances, has been discussed earlier, while the dubious honour of the name with the highest level of homonymy presently held in IRMNG goes to "Wagneria" with 14 separate instances (Figure 10).

It should also be noted that at specific level, certain slightly different epithet spellings are "deemed to be identical" under the zoological Code (Article 58) such as *caeruleus*/*coeruleus*/*ceruleus*, or *litoralis*/*littoralis*. These are, however, not reported as homonyms in the IRMNG species lists at this time, which is presently restricted to exactly matching epithets and associated generic names.

Homonyms at generic level remain the biggest source of potential confusion for both acquisition of IRMNG content (e.g., determining to what genus name instance to attach incoming species data) and for users wishing to resolve taxonomic names to a position in the taxonomic hierarchy, especially at specific level where the author of the genus (as opposed to the species name) will normally not be included. The ability to generate lists of homonymous names, or simply to check an existing name to determine whether or not any homonyms may exist, is also a useful feature for taxonomists and has been employed in a number of cases to date in order to discover previously unsuspected cases of homonymy at generic level in particular,

resulting in the proposal of replacement names as required (e.g., Ng and Low 2010; Zeidler 2017).

IRMNG CURRENT CLIENTS AND USE CASES

Web clients of IRMNG data over the past ten years have included a wide range of representatives of museums and herbaria, individual researchers, and members of the public from most countries in the world, as evidenced by accumulated logs of user searches by IP address and search term entered, plus email communications from specific users, typical enquiry rates via the web being in the order of tens to hundreds per day. A second, significant class of users comprises administrators of taxonomic information systems who prefer to receive a bulk download of IRMNG content (currently as data files, in future potentially also via a web service) for ingestion and re-use within their own systems. Once loaded there, re-usage via such "bulk clients" is of course supplementary to that recorded in user logs recording traffic via the IRMNG web interface.

The number of such bulk clients has grown over the past ten years and currently includes OBIS, GBIF (Döring 2017), the Atlas of Living Australia (ALA)[26], WoRMS, the Open Tree of Life project (OToL[27]; see also Hinchcliffe et al. 2015), the Global Names Index (GNI), and the Encyclopedia of Life (EoL)[28]. From 2017, the use of IRMNG data is also being investigated to potentially fill gaps in the present algal coverage of the Catalogue of Life[29]. A further capability, recently released at VLIZ, involves making IRMNG data accessible to machine-based query via dedicated web services, which opens up the potential to embed IRMNG queries as one element in a chain of machine-based reasoning: for example, a given query could seek habitat or extant/fossil status from IRMNG on a particular taxon before deciding whether to proceed further and extract additional information from a separate data source.

Specific uses of IRMNG include:

- Parse incoming or stored names of initially unknown taxonomic affinity, allocate to position in a taxonomic hierarchy based on their full species name when held, or the genus portion of the species name as applicable (also alerting to possible

---

[25] http://www.irmng.org/homonyms.php

[26] http://www.ala.org.au/faq/species-data/
[27] https://tree.opentreeoflife.org/about/open-tree-of-life
[28] http://eol.org/content_partners/676
[29] http://www.catalogueoflife.org/col/details/database/id/501

Homonyms

IRMNG homonym list - family level (96)

Please note, this is a provisional homonyms list from IRMNG on 2017-04-17, and is subject to change. It may include *nomina nuda*, rejected and otherwise non-current names, plus some published misspellings, in addition to homonyms that have been validly published. A small proportion of duplicate entries awaiting rationalisation may also be included.

Click on any family name for additional details, as currently available.

Acrolepidae
❌ Acrolepidae accepted as Glyphipterigidae (Misspelling of Acrolepiidae) · Animalia (Kingdom) > Arthropoda (Phylum) > Insecta (Class) > Lepidoptera (Order) > Acrolepidae (Family)
❌ Acrolepidae Aldinger, 1937 † accepted as Acrolepididae Aldinger, 1937 † · Animalia (Kingdom) > Chordata (Phylum) > Actinopterygii (Class) > Palaeonisciformes (Order) > Acrolepidae (Family)

Admetidae
❌ Admetidae Troschel, 1865 accepted as Cancellariidae Gray, 1853 · Animalia (Kingdom) > Mollusca (Phylum) > Gastropoda (Class) > Neogastropoda (Order) > Admetidae (Family)
❌ Admetidae Pocock, 1899 accepted as Phrynidae Blanchard, 1852 · Animalia (Kingdom) > Arthropoda (Phylum) > Arachnida (Class) > Amblypygi (Order) > Admetidae (Family)

Amphibolidae
Amphibolidae Hubendick, 1945 · Animalia (Kingdom) > Mollusca (Phylum) > Gastropoda (Class) > Basommatophora incertae sedis (Order) > Amphibolidae (Family)
❌ Amphibolidae Bertolani, 1981 accepted as Eohypsibiidae Bertolani & Kristensen, 1987 · Animalia (Kingdom) > Tardigrada (Phylum) > Eutardigrada (Class) > Parachela (Order) > Amphibolidae (Family)

Amphiporidae
Amphiporidae Rukhin, 1938 † · Animalia (Kingdom) > Porifera (Phylum) > Stromatoporoidea (Class) > Stromatoporida (Order) > Amphiporidae (Family)
Amphiporidae Oersted, 1844 · Animalia (Kingdom) > Nemertea (Phylum) > Enopla (Class) > Hoplonemertea (Order) > Amphiporidae (Family)

Aplodontidae
❌ Aplodontidae Brandt, 1855 accepted as Aplodontiidae Brandt, 1855 · Animalia (Kingdom) > Chordata (Phylum) > Mammalia (Class) > Rodentia (Order) > Aplodontidae (Family)
❌ Aplodontidae Kuroda, 1933 accepted as Modulidae P. Fischer, 1884 · Animalia (Kingdom) > Mollusca (Phylum) > Gastropoda (Class) > Caenogastropoda incertae sedis (Order) > Aplodontidae (Family)

Figure 8. Initial portion of the IRMNG family-level homonyms list as generated in May 2017; unaccepted names are indicated by the red circle enclosing an 'x', entirely extinct taxa by the dagger suffix (†). In the case of the two instances of family Amphiporidae within the same nomenclatural Code (one in Nemertea based on *Amphiporus* Ehrenberg, 1831, one in Porifera based on *Amphipora* Schulz, 1883), an application has been made (Özdikmen and Demir 2011) to remove the homonymy by emending the junior name (in Porifera) to Amphiporaidae; at the time of writing this case still awaits a decision from the relevant Commission (International Commission on Zoological Nomenclature 2017).

Figure 9. Initial portion of the "complete" IRMNG genera list, sorted alphabetically, showing the presence of 4 homonym pairs (8 names), as indicated by red outlines, within the first 33 names listed as at May 2017.

Figure 10. IRMNG holdings for genus name = "Wagneria" as at May 2017. Note: synonymy has been researched for many, but not all IRMNG name instances at the present time, therefore a subset of names is not yet flagged as synonym in the above list.

homonyms, i.e. a non-unique generic name within the IRMNG system): also referred to as a taxonomic name resolution service (see Discussion).

- Obtain near ("fuzzy") matches to an input name or names to cope with candidate misspelled names in incoming or presently stored data (a second component of a taxonomic name resolution service).

- Allocate incoming or stored names to categories i.e. extant, fossil or both; marine, nonmarine or both, for filtering and/or sorting purposes as desired.

- Generate top-down views of "all life" to at least generic level by traversing a taxonomic tree of kingdom through genus (and species where available), for comparison/integration with equivalent data from other sources.

- Generate lists of taxa and names on demand based on any characteristic held in the database (for example name begins with…, authority contains…), also generate lists of duplicate names (homonyms) as required.

- Provide cross walks / deep links to data held remotely in other systems (at this time WoRMS, ION, and BioNames) by holding those systems' taxon or name identifiers within relevant IRMNG records.

A schematic view of IRMNG data flows including sources, editor actions, and the range and nature of current clients as indicated above is represented in Figure 11.

IRMNG EDITING

Since its inception in 2006, IRMNG data compilation and editing has been the responsibility of OBIS Australia, whose personnel have been assigned privileges at the time to edit all aspects of content in their designated taxonomic group/s, either as direct operations on the database or on offline copies of relevant data files that have then been uploaded to the live system. From 2016 onwards, a new web-based edit system has been constructed which permits authorized editors to similarly alter any aspects of taxa under their control including the creation and quarantining of records, correcting spelling errors, adding, changing, or deleting attributes, sources and links, moving child records to a different parent, and so on. Responsibility for allocation of edit privileges to relevant persons for the future now resides with the VLIZ Data Management Team who will be managing this aspect in tandem with equivalent procedures for WoRMS (for more details see Discussion).

IRMNG DATA AVAILABILITY

Data dumps of IRMNG content in DwC-A and/or native database format have been made available to clients on request since 2007, with a new dedicated download location created at the VLIZ instance via the URL[30], from which (at time of writing) the last three "snapshot" versions of IRMNG are available dating from January 2013, January 2014, and April 2017. Meanwhile, the master version of IRMNG content is always accessible via the web and may contain updates that post-date any particular snapshot, plus in addition some supplementary information (principally sources for assertions used, together with the searchable literature module) not included in the Darwin Core data file(s). Historic data dumps from IRMNG are also available via other locations; for example, at time of writing the 2013 version is searchable via the Global Names Resolver[31] and the 2014 version is searchable via a copy hosted at GBIF[32], plus an archived version is accessible via the holdings of the Open Tree of Life[33].

IRMNG data are released without any IRMNG-issued copyright assertion, although unfortunately the same is not true for some of its constituent sources. WoRMS data is currently CC-BY (data are freely available for re-use but attribution is required), the Catalogue of Life and Index Fungorum declare their content to be available for re-use by non-commercial users only[34,35], while (for example) ION, Zoological Record and AlgaeBase assert that their content "may not be downloaded or replicated by any means" without appropriate permission[36,37].

Nevertheless, as argued by Patterson et al. (2014), a particular taxonomic name, its cited authority, taxonomic position, status and so on are simply facts or opinions sourced from the primary scientific literature and should not therefore be copyrightable by any downstream compilations. An exception may exist for comments added by record editors but arguably these should also be reproducible under "fair use", with appropriate attribution, as is the case with such comments from other sources reproduced in IRMNG. For so long as the IPR

---

[30] http://www.irmng.org/download.php
[31] http://resolver.globalnames.org/
[32] http://www.gbif.org/dataset/0938172b-2086-439c-a1dd-c21cb0109ed5
[33] http://purl.org/opentree/ott/ott2.8/inputs/IRMNG_DWC-2014-01-30.zip
[34] http://www.catalogueoflife.org/content/terms-use
[35] http://www.indexfungorum.org/Names/IndexFungorumPartnership.htm
[36] http://organismnames.com/terms.htm
[37] http://www.algaebase.org/copyright/
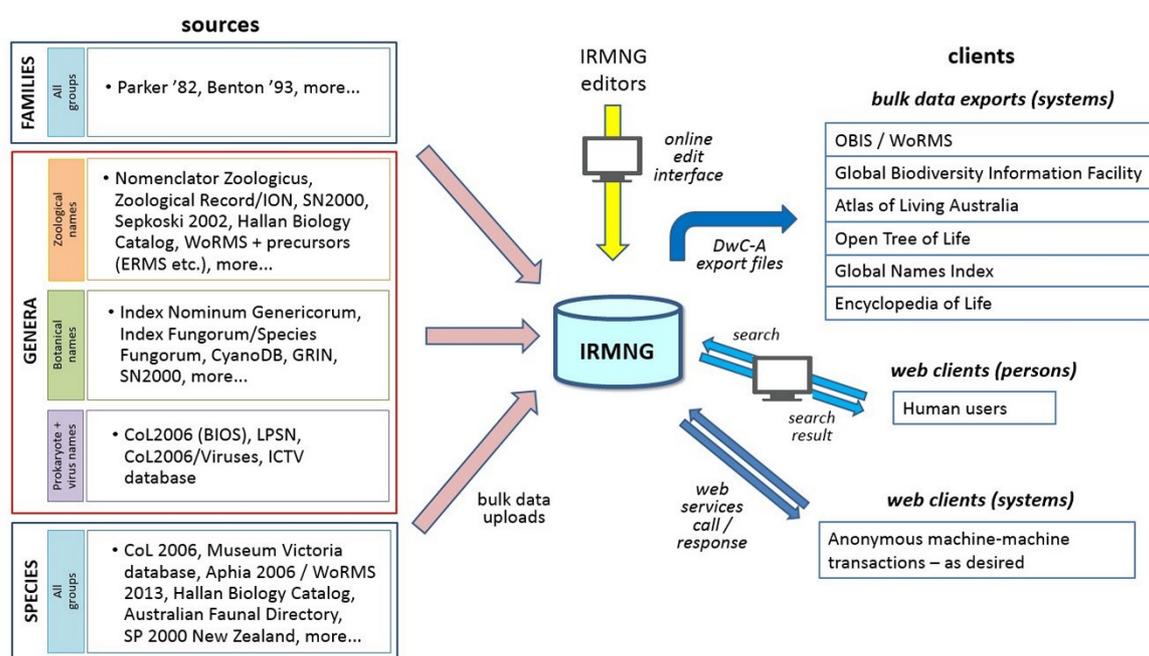
Figure 11. Schematic overview of IRMNG data flows. A wide variety of specialized and more general sources are used to populate the IRMNG database. All name-related sources go through a number of pre-processing steps prior to upload, including comparison with already available IRMNG data and assigning of the relevant flags. IRMNG content can also be added or upgraded remotely by authorized editors using an online edit interface. Exports of IRMNG content to external systems are arranged either through Darwin Core Archive (DwC-A) files or through web services, while human users can enter search queries and be returned relevant results via the IRMNG web portal.

situation remains untested, IRMNG data downloads are presently accompanied by a statement "IRMNG data to specific level incorporates some content from the Catalogue of Life, the World Register of Marine Species (WoRMS) and other providers and may be subject to their respective terms of use" and are not accompanied by a free-reuse license such as Creative Commons CC0[38]; however, it is to be hoped that this may change in the future. IRMNG does not require attribution for its data (which in any case originate almost exclusively from other sources) although an acknowledgement is requested if users find its services of value in their work.

DISCUSSION
Other taxonomic compilations

IRMNG is presently unique in that it aspires to completeness, at generic level at least, for "all life," within a coherent hierarchical data structure which attempts to conform to current taxonomic opinion, to the degree that present resources permit. In this respect it differs from strict nomenclatural compilations such as *Nomenclator Zoologicus*, *Index Nominum Genericorum*, the International Plant Name Index (IPNI)[39], ION and others which, in addition to their self-imposed taxonomic coverage limits, are concerned in the first instance with the date, authorship, and place of publication of scientific names (facts) and less on their valid name/synonym status and current taxonomic placement (both opinions, also subject to change through time). Compilations that do share these goals include, e.g., ITIS, the Catalogue of Life (chiefly for extant taxa) and the Paleobiology Database (for fossils), as well as numerous more selective compendia such as WoRMS. Conceptually, IRMNG presently fills a gap that eventually should be occupied by content from a combination of the Catalogue of Life and the Paleobiology Database, if these compilations were complete and if the Catalogue of Life were extended to hold information on genera, and thus it is pertinent to assess their present levels of completeness with respect to IRMNG, in particular at generic level. Some relevant comparisons with these and other available sources are given in Table 3.

From the data presented in Table 3 it can be seen that IRMNG is currently the most extensive, vetted compendium of genus names across all groups and in addition, contains the extant/fossil and marine/nonmarine status flags of interest to a range of users for over 80% of its entries at generic level. Only GNI (not utilized as an IRMNG source to date) is likely to contain a potentially useful component of genus names not yet in IRMNG, but the effort to extract these would most likely be considerable, bearing in mind that GNI contains a range of misspelled, malformed, or even non-names (e.g., vernacular names harvested by mistake) as well as multiple variants of existing names.

In addition, since the GNI data structure does not distinguish genus names from uninomials at other ranks, each candidate "new" name would have to be researched manually to see if it were even a genus before considering further for aspects of interest to IRMNG such as correct cited authority, taxonomic status, current placement and more. On the other hand, potential sources of additional species names include such compilations as ION, Catalogue of Life, the PaleoBiology Database, IPNI and more, from which the expert vetted component(s) would certainly be useful to add to IRMNG at some point in the future.

Large scale aggregators such as Global Names, GBIF, Encyclopedia of Life, and Open Tree of Life are a special case in that they are consumers of IRMNG content (among others) which they then integrate to produce "supersets" including IRMNG data in combination with other information (e.g., see Döring 2017; Rees and Cranston 2017), in the same way that IRMNG is itself a data aggregator. This then generates a question as to whether such aggregators could, or should, then be used as a preferred alternative to IRMNG for general taxonomic name resolution. Positives of this approach include the fact that a user will encounter non-IRMNG as well as IRMNG-sourced data which can be useful to fill data gaps, especially at specific level.

On the other hand, the IRMNG content in such systems will generally be a time-stamped snapshot, and "live" IRMNG data may have been adjusted/improved in the intervening time; IRMNG may have richer content on a given taxonomic name than is preserved in the version made available elsewhere; and IRMNG may also be searchable via ways not offered in the data aggregator in question. By the same logic,

---

[38] https://wiki.creativecommons.org/wiki/CC0/
[39] http://www.ipni.org/

Table 3. IRMNG holdings and attributes covered compared with major other relevant compilations as at end 2016.

Data Source	Scope	No. of genus names held	No. of species names held	Genus authorities held?	Family allocation held? (for subsidiary ranks)	Accepted name/synonym status held?	Extant/fossil status held?	Marine/non-marine status held?	Remarks
IRMNG	All groups, extant + fossil, kingdoms to species	487,000	1.78 million	yes	yes (genera 78%, species >95%)	yes (genera 67%, species >95%)	yes (genera 85%, species >95%)	yes (genera 88%, species >90%)	Publication details (as microcitations) held for many genera, mostly in zoology.
Catalogue of Life, 2016 version	85% of all extant groups, small number of fossils, kingdoms to subspecies	193,000	3.1 million	no	yes	species: yes; genera: no	yes	no	Taxonomic coverage of extant groups currently incomplete (e.g., all algae, some arthropods missing). No publication details held for any names.
Global Names Index	All groups, extant + fossil, all ranks	400,000-600,000?	c. 16 million	yes	no	no	no	no	No exact genus count available. Includes many duplicated names with minor authority variants, and misspellings. Unique binomial name count is 6.8 million if authorities are ignored (D. Mozzherin pers. comm. 2016).
ION	Zoological names, extant + fossil, all ranks	300,000-400,000?	c. 3 million	yes	for subset only (<30%)	no	no	no	No exact genus count available. Many names held multiple times (authority variants not deduplicated), misspellings not corrected.
PaleoBiology Database, 2016 version	Zoological plus botanical names, almost all fossil, all ranks	83,000	159,000	yes	yes	yes	yes	yes	No detailed information on website, values are from 2016 data dump (Australian Research Council 2016). Detailed geologic and environmental information is held as available; many names have full bibliographic citations.
Nomenclator Zoologicus	Zoological names, extant + fossil, genera and subgenera only, cut-off date 2004	357,000	0	yes	no	no	yes	no	Higher taxonomic allocation typically given to phylum (or class / order) only, according to group. Fossil indication included for many but not all applicable names.
Index Nominum Genericorum	Botanical names, extant + fossil, genera only, 1753-2012 (+?)	69,000	0	yes	yes	no	yes	no	Higher taxonomic allocation typically given to family, but outdated in some cases.

IPNI	Botanical names excluding algae, fungi and bryophytes, extant only, all ranks	30,000+?	0.8-1 million?	yes	yes	no	yes (all presumed extant)	no	Over 1.3 million names at all ranks, including infraspecies. Generic names will largely overlap coverage of equivalent groups in *Index Nominum Genericorum*.
ITIS	Most groups, all extant, all ranks	50,000-150,000?	526,000	yes	yes	yes	yes (all presumed extant)	no	708,000 names at all ranks as at December 2015, including 526,000 species binomials (Guala 2016).
WoRMS	Marine taxa (plus some nonmarine), extant + some fossil, all ranks	47,000	470,000	yes	yes	yes	yes	yes	Fossil taxa are included for selected groups only.

IRMNG data may not always represent the most current version of its own externally sourced content, and it is also recommended to visit those sites too for their most current content. That aspect aside, IRMNG aspires to provide a single point of entry to a comprehensive overview of "all life", particularly to generic level, with added, machine-readable basic extant/fossil and habitat indicators, as well as associated original publication information, taxonomic comments and more, that may not be readily accessible elsewhere in such an integrated form.

IRMNG as a research entity

A compilation of genus names through time such as IRMNG offers a rich source of content for further study, such as the numbers of genera per taxonomic group published through time, the lexical character and variation of scientific names, overall contributions by specific workers, reporting the size of different taxonomic groups, ratios of valid names to synonyms once the latter are more completely annotated, compilation of lists of within-Code and between-Code homonyms, and much more. Such topics are largely outside the scope of the present paper but may be discussed elsewhere in due course. In this respect IRMNG can function as a literary corpus, which is a standardized collection of text upon which specific linguistic investigations can be performed and results reported in a repeatable fashion (Biber et al. 1998). A further valuable use of the collection of names assembled within IRMNG has been in providing a reference suite of data to support the development and extensive testing of algorithms for name matching, including both Taxamatch and some subsidiary methods, as reported in more detail in Rees (2014). Representing as it does the current largest, deduplicated and preliminarily vetted set of genus names so far compiled across all taxonomic domains, virtually all with an associated authority and year, IRMNG is presently unique in offering potential insights into the nature of such names as well as the history of contributions by different scientists through time.

IRMNG as a source of "marine" content

IRMNG taxon names currently flagged "marine = yes" that are not presently held in WoRMS represent a potential source of additional content to the latter system. A workflow has been commenced whereby such names within groups of interest to WoRMS can be passed to relevant WoRMS experts for scrutiny as needed and then added to WoRMS holdings once approved to do so. A separate, but related issue, is to compare environmental flags as presently held in the two systems with an emphasis on upgrading these, in particular within WoRMS. This work (recommended by the WoRMS Steering Committee, Dec. 2015) has commenced, and in this context the flags already present in IRMNG for many taxa have already proved a valuable resource to assist both the VLIZ Data Management Team and WoRMS taxonomic editors.

IRMNG as a taxonomic name resolution service

With its associated tools (name parsing and fuzzy matching) accessing a reference database of hierarchically arranged scientific names, the web "search" function provided to IRMNG users since 2006 can be viewed as one of the earliest implementations of a taxonomic name resolution service, which has been characterized as "a service that corrects variant and erroneous spellings, disambiguates homonyms by means of higher taxonomic filtering, and updates synonyms with reference to authoritative taxonomic sources" (Boyle et al. 2013). Such a service (currently in version 4.0[40]) is also offered by the iPlant taxonomic name resolution service (TNRS) described by Boyle et al., which has been operational since 2011. This service offers many of the same features as IRMNG but differs in scope, since it is currently set up to access five databases of land plants only (i.e., extant bryophytes through angiosperms) and thus currently excludes algal, fungal, protozoan, and animal names, as well as all fossils. There are some minor conceptual differences between the iPlant TNRS and IRMNG in that in the iPlant TNRS, fuzzy matching is extended to family level and name parsing is performed by a dedicated module developed elsewhere, the Global Names Index Scientific Names Parser[41]. Most significantly, in the case of the iPlant TNRS, all of the reference datasets searched exist as separate external resources whose curation and maintenance are not the responsibility of the iPlant development team.

IRMNG sustainability and future activities

As with any biodiversity project intended to last beyond just a few years, succession planning

---

and the sustainability of the project are important considerations, especially as the project lifetime extends (Costello et al. 2014). Initially the project was developed to address the taxonomic and associated trait requirements of OBIS and progressed in accordance with that project's direction (e.g., UNESCO 2012), as a contribution from the OBIS Australia Regional Node which was at that time (and is still) located at CSIRO in Australia.

When it was realised that this contribution was potentially coming to an end, a succession plan was introduced which involved transferring the data and internal relationships to the servers at VLIZ where they could be sustained by that institution into the future in tandem with that organisation's pre-existing commitment to maintain WoRMS and associated taxonomic databases[42]. From 2016, the VLIZ Data Management Team (DMT) have assumed responsibility for ongoing maintenance and development of both IRMNG and the IT platform that supports it, addressing most of the concerns expressed in Costello et al. (2014) with the strictly IT-related costs essentially already covered under the general operation of existing VLIZ data systems and therefore not requiring separate resourcing. With this move, IRMNG has also evolved from a system largely developed and maintained principally by a single author to one supported by a larger team and with the potential for distributed content maintenance and enhancement in the future, again a key recommendation of the combined expertise represented by the authors contributing to Costello et al. (2014).

In its new location as part of the VLIZ "family" of taxonomic data systems, a number of new as well as continuing activities are envisaged. First, the public web presence and search-based functions are now available at the new location, modified to reflect and include functionality already developed for WoRMS and associated data systems that also run on the Aphia platform. Second, export routines for IRMNG data are being continued and new versions will be made available to present and also potential new IRMNG clients. Third, IRMNG data will be made available to external automated clients via SOAP/WSDL-based web services as already offered for other VLIZ-hosted databases (SOAP: Simple Object Access

Protocol; WSDL: Web Services Description Language; for additional explanation see e.g., Weerawarana et al. 2005). Fourth, co-location of the IRMNG and WoRMS/OBIS data systems is facilitating closer linkages between these projects, as well as with the Taxonomic Backbone for the European LifeWatch project which is also being developed at VLIZ (Dekeyzer et al. 2014, see also the LifeWatch page at VLIZ[43] and the LifeWatch "Taxonomic Information" page[44]). Fifth, the remote editing interfaces already developed for use by WoRMS editors have been adapted for use with IRMNG and are already operational and in use by relevant accredited users. Lastly, VLIZ has also expressed a desire to develop a network of remote editors for IRMNG along the lines of that already in place for WoRMS, a process which is anticipated to be commenced over the period 2017-2018. It is also intended to avoid duplication of effort so far as is possible, in other words if a WoRMS editor already exists for a predominantly marine group, WoRMS would be the natural continued vehicle to hold this information from which it can be ported to IRMNG in due course without the requirement for additional data entry.

On top of additional functionality and linkages for the new version of IRMNG, the requirement for ongoing maintenance and content enhancement will continue. This can conceptually be divided into three areas: (1) ongoing addition of new names as these are published—the previously noted approx. 2,500 names annually at generic rank (plus new families as well), in addition to between 15,000 and 20,000 new species names per year, to the extent that the latter are considered desirable to be held; (2) "catch-up" acquisition of legacy names not yet held—particularly additional species if these are desired—using available sources such as more recent editions of the Catalogue of Life, PaleoBiology Database, and others; and (3) upgrade of content already held, with respect to the present data limitations as discussed earlier.

One aspect which should receive particular attention is a review of the present IRMNG classification above family level, which (with the exception of specific groups as detailed in Appendix 1) retains some inconsistencies on

---

account of being imported from a range of different sources over time, in many cases because no "standard" higher taxonomy existed at that time. With the emergence of some recent standardized treatments such as Zhang (2011, 2013), Ruggiero (2014), and Ruggiero et al. (2015), as well as the ongoing contributions by the Angiosperm Phylogeny Group already incorporated, the opportunity exists to review and standardize the IRMNG classification to a higher degree than previously implemented, and thereby facilitate better interoperability and data exchange with other systems that utilize the same classifications. To date, prioritization and resourcing of the various areas of activity indicated above have been the responsibility of the first author in the main, aided by in-kind contributions from CSIRO plus grants from key IRMNG clients including OBIS, GBIF and the Atlas of Living Australia. From 2016 onwards, future directions and resourcing for continued IRMNG development will be determined by the VLIZ data management team (which includes the remaining authors of this paper) in liaison with their network of other interested parties including OBIS, WoRMS and its editors, Catalogue of Life and more.

CONCLUDING REMARKS

When IRMNG was commenced in 2006, its main role was to provide a means for OBIS and related data aggregation projects to manage incoming taxonomic data in a consistent manner, and to distinguish marine from nonmarine, and extant from fossil taxa. As an accessory function, the system has also served as an effective taxonomic name resolution service for these projects and a range of other users over the succeeding ten years. During that time, there has been a growth in the availability and/or completeness of other systems dealing with species data in particular, such as the Catalogue of Life annual editions, the Paleobiology database, the World Register of Marine Species, and The Plant List, which are currently more complete than IRMNG for species although less so for genera. The existence of "super aggregators" such as the Global Biodiversity Information Facility, the Global Names Index and Global Names Resolver, the Encyclopedia of Life, and the Open Tree of Life also permits such projects to claim more comprehensive coverage than IRMNG in the simple metric of "numbers of names held", although a number of these ingest IRMNG data and therefore rely to some degree

on IRMNG to continue to acquire and supply them with new content.

In its location at VLIZ from 2016 onwards, and with new arrangements for governance, data sharing and distributed editing, the method of operation for IRMNG is certain to evolve and change in some respects as well as responding to changes in the overall Biodiversity Informatics landscape. This review of the first ten years of IRMNG's operation serves as a document of processes and activity in content building to date and provides a baseline against which future progress can be measured, and in addition may potentially provide some ideas and discussion points of value to other biodiversity projects that operate in a similar area, both now or in the future.

AUTHOR CONTRIBUTIONS

Tony Rees established the original IRMNG data compilation, designed, constructed, and populated the initial IRMNG data tables at CSIRO and created the associated web application for data entry and retrieval. Leen Vandepitte collaborated with Tony to manage the transfer of IRMNG data to VLIZ, while Wim Decock and Bart Vanhoorne carried out the technical operations of data mapping between the CSIRO and Aphia systems, migration to the VLIZ servers and construction of the web interfaces for the new version of IRMNG hosted at VLIZ, and serve as administrators for the system from 2016 onwards. All authors contributed to pre-release testing, release, and ongoing improvement of the new version of IRMNG at VLIZ during 2016. Tony Rees wrote the initial version of this manuscript which was then developed further with the benefit of contributions from the other authors.

LITERATURE CITED

Agassiz, L. 1848. *Nomenclatoris zoologici index universalis: continens nomina systematica classium, ordinum, familiarum et generum animalium omnium, tam viventium quam fossilium, secundum ordinem alphabeticum unicum disposita, adjectis homonymiis plantarum, nec non variis adnotationibus et emendationibus.* Soloduri: Sumptibus Jent et Gassman.

Allkin, R., R.J. White, and P.J. Winfield. 1992. Handling the taxonomic structure of biological data. Mathl. Comput. Modelling 16(6/7):1-9.

Alroy, J. 2002. How many named species are valid? Proc. Natl. Acad. Sci. USA 99:3706-3711.

Andrews, H.N. 1970. Index of Generic Names of Fossil Plants, 1820-1965. Geological Survey Bulletin 1300. US Government Printing Office, Washington; plus supplements (Geological Survey Bulletin 1396: Index of Generic Names of Fossil Plants 1966-1973 by A. M. Blazer, 1975, and Geological Survey Bulletin 1517: Index of Generic Names of Fossil Plants 1974-1978 by A. D. Watt, 1982).

Angiosperm Phylogeny Group. 2009. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. Bot. J. Linn. Soc. 161:105-121.

Angiosperm Phylogeny Group. 2016. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. Bot. J. Linn. Soc. 181:1-20.

Appeltans, W., S.T. Ahyong, G. Anderson, M.V. Angel, T. Artois, et al. 2012. The magnitude of global marine species diversity. Curr. Biol. 22:2189-2202.

Australian Research Council. 2016. The Paleobiology Database (July 5 2016 version), hosted by the GBIF Secretariat. Online web resource, available at http://www.gbif.org/dataset/c33ce2f2-c3cc-43a5-a380-fe4526d63650.

Benton, M.J., ed. 1993. The Fossil Record 2. Chapman & Hall. London.

Benton, M.J. 2008. How to find a dinosaur, and the role of synonymy in biodiversity studies. Paleobiology 34:516-533.

Biber, D., S. Conrad, and R. Reppen. 1998. Corpus Linguistics—Investigating Language Structure and Use. Cambridge University Press, Cambridge.

Bisby, F.A., M.A. Ruggiero, Y.R. Roskov, M. Cachuela-Palacio, S.W. Kimani, P.M. Kirk, A. Soulier-Perkins, and J. van Hertum, eds. 2006. Species 2000 & ITIS Catalogue of Life: 2006 Annual Checklist. CD-ROM; Species 2000, Reading.

Bouchard, P., Y. Bousquet, A. Davies, M. Alonso-Zarazaga, J. Lawrence, C. Lyal, A. Newton, C. Reid, M. Schmitt, A. Slipinski, and A. Smith. 2011. Family-group names in Coleoptera (Insecta). ZooKeys. 88:1-972

Bouchet, P., and J.-P. Rocroi. 2005. Classification and nomenclator of gastropod families. Malacologia 47:1-397.

Boyle, B., N. Hopkins, Z. Lu, J.A. Raygoza Garay, D. Mozzherin, T. Rees, N. Matasci, M.L. Narro, W.H. Piel, S.J. Mckay, S. Lowry, C. Freeland, R.K. Peet, and B.J. Enquist. 2013. The taxonomic name resolution service: an online tool for automated standardization of plant names. BMC Bioinformatics 14:16.

Chapman, A.D. 2009. Numbers of Living Species in Australia and the World. Australian Government,

Department of the Environment, Canberra, 2nd edition.

Costello, M.J. 2000. Developing species information systems: the European Register of Marine Species (ERMS). Oceanography 13(3):48-55.

Costello, M.J., W. Appeltans, N. Bailly, W.G. Berendsohn, Y. de Jong, M. Edwards, R. Froese, F. Huettmann, W. Los, J. Mees, H. Segers, and F.A. Bisby. 2014. Strategies for the sustainability of online open-access biodiversity databases. Biol. Conserv. 173:155-165.

Costello, M.J., C. Emblow, and R.J. White, eds. 2001. European register of marine species: a check-list of the marine species in Europe and a bibliography of guides to their identification. Collection Patrimoines Naturels, 50. Muséum national d'Histoire naturelle, Paris.

De Grave, S., N.D. Pentcheff, S.T. Ahyong, T.-Y. Chan, K.A. Crandall, P.C. Dworschak, D.L. Felder, R.M. Feldmann, C.H.J.M. Fransen, L.Y.D. Goulding, R. Lemaitre, M.E.Y. Low, J.W. Martin, P.K.L. Ng, C.E. Schweitzer, S.H. Tan, D. Tshudy, and R. Wetzer. 2009. A classification of living and fossil genera of decapod crustaceans. Raffles Bull. Zool., Suppl. 21:1-109.

Dekeyzer, S., L. Vandepitte, S. Claus, K. Deneudt, and F. Hernandez. 2014. The LifeWatch taxonomic backbone: supporting the marine biodiversity and ecosystem functioning research community [abstract]. P. 106 in A Variety of Interactions in the Marine Environment. Abstracts volume from 49th European Marine Biology Symposium, September 8-12, 2014. St. Petersburg, Russia.

Döring, M. 2017. GBIF backbone—February 2017 update. GBIF Developer Blog, available at http://gbif.blogspot.com.au/2017/02/gbif-backbone-february-2017-update.html.

Eschmeyer, W.N. 2000-current. Catalog of Fishes Online Database. Online web resource, available at http://researcharchive.calacademy.org/research/ichthyology/catalog/fishcatmain.asp.

Euzéby, J.P., ed. plus successors. 1997-current. List of Prokaryotic Names with Standing in Nomenclature (originally: List of Bacterial Names with Standing in Nomenclature). Online web resource, available at http://www.bacterio.net. Edited by J. P. Euzéby, 1997-2013, A. C. Parte from 2013.

Farr, E.R., J.A. Leussink, and F.A. Stafleu, eds. 1979. *Index Nominum Genericorum (Plantarum)*. Scheltema & Holkema, Utrecht, and W. Junk, The Hague. 3 volumes.

Gordon, D.P., ed. 2009-2011. New Zealand Inventory of Biodiversity. Volume 1: Kingdom Animalia: Radiata, Lopotrochozoa, Deuterostomia; Volume 2: Kingdom Animalia: Chaetognatha, Ecdysozoa, Ichnofossils; Volume 3: Kingdoms Bacteria, Protozoa, Chromista, Plantae, Fungi. Canterbury University Press, Canterbury, New Zealand.

Guala, G.F. 2016. The importance of species name synonyms in literature searches. PLoS ONE 11(9):e0162648.

Hinchliff, C.E., S.A. Smith, J.F. Allman, J.G. Burleigh, R. Chaudhary, L.M. Coghill, K.A. Crandall, J. Deng, B.T. Drew, R. Gazis, K. Gude, D.S. Hibbett, L.A. Katz, H.D. Laughinghouse, E.J. McTavish, P.E. Midford, C.L. Owen, R.H. Ree, J.A. Rees, D.E. Soltis, T. Williams, and K.A. Cranston. 2015. Synthesis of phylogeny and taxonomy into a comprehensive tree of life. Proc. Nat. Acad. Sci. USA 112:12764-12769.

Hooker, J.D., and B.D. Jackson. 1895. *Index Kewensis plantarum phanerogamarum nomina et synonyma omnium generum et specierum a linnaeo usque ad annum MDCCCLXXXV...* Oxford University Press, Oxford.

International Commission on Zoological Nomenclature. 1999. International Code of Zoological Nomenclature, 4th edition. International Trust for Zoological Nomenclature, London.

International Commission on Zoological Nomenclature. 2017. Open cases. Available online at http://iczn.org/content/list-open-cases, date of query 25 May 2017.

International Committee on Taxonomy of Viruses (ICTV). 1971-current. Report(s) on Virus Classification: includes 7th Report (1999), 8th Report (2005), 9th Report (2009).

Kirk, P., compiler. 2001-current. *Index Fungorum*. Online web resource, available at http://www.indexfungorum.org/.

Lapage, S.P., P.H.A. Sneath, E.F. Lessel, V.B.D. Skerman, H.P.R. Seeliger, and W.A. Clark, eds. 1992. International Code of Nomenclature of Bacteria: Bacteriological Code, 1990 Revision. ASM Press, Washington (DC), U.S.A.

Martin, J.W., and G.E. Davis. 2001. An Updated Classification of the Recent Crustacea. Science Series 39, Natural History Museum of Los Angeles County.

McNeill, J., F.R. Barrie, W.R. Buck, V. Demoulin, W. Greuter, D.L. Hawksworth, P.S. Herendeen, S. Knapp, K. Marhold, J. Prado, W.F. Prud'homme van Reine, G.F. Smith, J.H. Wiersema, and N.J. Turland. 2012. International Code of Nomenclature for algae, fungi, and plants (Melbourne Code). Regnum Vegetabile 154. Koeltz Scientific Books.

Mullins, G.L., ed. 2007. The phytoPal Taxonomic Database (taxon list), version 02 February 2007. 1132 pp. Originally downloaded from http://www.le.ac.uk/geology/glm2/phytopal/taxa.pdf (no longer available from that source).

Neave, S.A., ed. 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 (plus succeeding volumes to end 1994). Zoological Society of London, London. 9 volumes.

Ng, P.K.L., and M.E.Y. Low. 2010. On the generic nomenclature of nine brachyuran names, with four replacement names and two *nomina protecta* (Crustacea: Decapoda). Zootaxa 2489:34-46.

Özdikmen, H., and Demir, H. 2011. Case 3540 AMPHIPORIDAE Rukhin, 1938 (Porifera, Stromatoporata, Amphiporida): proposed emendation to AMPHIPORAIDAE to remove homonymy with AMPHIPORIDAE McIntosh, 1873 (Nemertea, Hoplonemertea). Bull. Zool. Nomenclature 68:167-169.

Parker, C.T., G.M. Garrity, and B.J. Tindall. 2015. International Code of Nomenclature of Prokaryotes. Int. J. Syst. Evol. Microbiol. November 2015.

Parker, S.P., ed. 1982. Synopsis and Classification of Living Organisms. McGraw-Hill Book Company, New York. 2 volumes.

Patterson, D.J., J. Cooper, P.M. Kirk, R.L. Pyle, and D.P. Remsen. 2010. Names are key to the big new biology. Trends Ecol. Evol. 25:686-691.

Patterson, D.J., W. Egloff, D. Agosti, D. Eades, N. Franz, G. Hagedorn, J.A. Rees, and D.P. Remsen. 2014. Scientific names of organisms: attribution, rights, and licensing. BMC Res. Notes 7:79.

Patterson, D., D. Mozzherin, D.P. Shorthouse, and A. Thessen. 2016. Challenges with using names to link digital biodiversity information. Biodiv. Data J. 4:e8080.

Prothero, D.R. 2013. Bringing Fossils to Life: An Introduction to Paleobiology. 3rd Edition. Columbia University Press, New York.

Raup, D.M. 1986. Biological extinction in Earth history. Science 231:1528-1533.

Rees, J.A., and Cranston, K. 2017. Automated assembly of a reference taxonomy for phylogenetic data synthesis. Biodiversity Data Journal 5:e12581.

Rees, T., 2008a. Applications of fuzzy (approximate string) matching in taxonomic database searches, with an example multi-tiered approach. P. 11-13 in Worcester, T., L. Bajona, and B. Branton, eds. Proceedings of a Conference on Ocean Biodiversity Informatics (2-4 October 2007). Bedford Institute of Oceanography, Dartmouth, Nova Scotia.

Rees, T. 2008b. IRMNG—The Interim Register of Marine and Nonmarine Genera [abstract]. P. 72 in Weitzman, A. and L. Belbin, eds. Proceedings of TDWG (2008), Fremantle, Australia. Biodiversity Information Standards (TDWG) and Missouri Botanical Garden.

Rees, T. 2014. Taxamatch, an algorithm for near ('fuzzy') matching of scientific names in taxonomic databases. PLoS ONE 9(9):e107510.

Rees, T., and Y. Zhang. 2007. Evolving concepts in the architecture and functionality of OBIS, the Ocean Biogeographic Information System. Pp. 167-176 in Vanden Berghe, E. et al., eds. Proceedings Ocean Biodiversity Informatics: International Conference on Marine Biodiversity Data Management, Hamburg, Germany 29 November to 1 December, 2004. VLIZ Special Publication 37.

Ruggiero, M.A. 2014. Families of All Living Organisms, Version 2.0.a.15, (4/26/14). Expert Solutions International, LLC, Reston, VA, USA. Tabulated data file (as "GGI Family Data") available via the Global Genome Initiative (GGI) Knowledge Portal, http://ggi.eol.org/downloads.

Ruggiero, M.A., D.P. Gordon, T.M. Orrell, N. Bailly, T. Bourgoin, R.C. Brusca, T. Cavalier-Smith, M.D. Guiry, and P.M. Kirk, 2015. A higher level classification of all living organisms. PLoS ONE 10(4):e0119248.

Schultz, P.W., and M. Huber. 2013. Revision of the worldwide Recent Pinnidae and some remarks of fossil European Pinnidae. Acta Conch. 13:1-164.

Sepkoski. J.J., Jr. 2002. A compendium of fossil marine animal genera. Bull. Amer. Paleontol. 363:1-560.

Sherborn, C.D. 1902-1933. *Index animalium; sive, index nominum quae ab A. D. MDCCLVIII generibus et speciebus animalium imposita sunt, societatibus eruditorum adiuvantibus*. Cambridge, U.K.

UNESCO. 2012. IODE Steering Group for OBIS (SG-OBIS), Second Session, 19-21 November 2012. Reports of Meetings of Experts and Equivalent Bodies, (English), UNESCO.

Vandepitte, L., B. Vanhoorne, W. Decock, S. Dekeyzer, A.T. Verbeeck, L. Bovit, F. Hernandez, and J. Mees. 2015. How Aphia—the platform behind several online and taxonomically oriented databases—can serve both the taxonomic community and the field of biodiversity informatics. J. Mar. Sci. Eng. 3:1448-1473.

Weerawarana, S., F. Curbera, F. Leymann, T. Storey, and D.F. Ferguson. 2005. Web Services Platform Architecture: SOAP, WSDL, WS-Policy, WS-Addressing, WS-BPEL, WS-Reliable Messaging and More. Prentice Hall PTR, Upper Saddle River, NJ, USA.

Willis, J.C. 1973. A Dictionary of the Flowering Plants and Ferns. Cambridge University Press, Cambridge, U.K., 8th edition.

WoRMS Editorial Board. 2007-current. World Register of Marine Species. Available from http://www.marinespecies.org at VLIZ.

Zhang, Z.-Q., ed. 2011. Animal biodiversity: an outline of higher-level classification and survey of taxonomic richness. Zootaxa 3148:1-237.

Zhang, Z.-Q., ed. 2013. Animal biodiversity: an outline of higher-level classification and survey of taxonomic richness (addenda 2013). Zootaxa 3703:1-82.

Zeidler, W. 2017. Validation of the replacement name *Eusceliotes* Stebbing, 1888 for the pelagic hyperiidean amphipod genus *Euscelus* Claus, 1879 (Crustacea: Amphipoda: Hyperiidea: Parascelidae), preoccupied by *Euscelus* Schoenherr, 1833 (Insecta: Coleoptera: Attelabidae). Zootaxa 4254:377-378.

APPENDIX 1: SIGNIFICANT SOURCES USED IN IRMNG DATA COMPILATION, 2006-2016

Date added	Source and web location (as applicable)	Content added	Remarks
1/08/2006	Print work: "Synopsis and Classification of Living Organisms" (Parker 1982), taxonomic index section digitized by D. Palomares and team (SeaLifeBase) as a contribution to IRMNG	6,900 extant families	Contains valid family names only, with no associated authorities. A small number of groups not included.
20/09/2006	*Systema Naturae* 2000 (SN2000) courtesy Dr. S. Brands, compiler (2006 version)[45]	234 additional families, 111,000 genus names	Mix of data from a range of original sources, including 62,000 "verified" and 49,000 "unverified" genus names, most with present family allocation. Species names not requested.
20/09/2006	Catalogue of Life, 2006 edition (Bisby et al. 2006)	2,100 additional families, 36,000 additional extant genus names plus 1,282,000 extant species names including synonyms (including 120,000 later deprecated as duplicates)	Includes numerous groups, viruses to vertebrates although holdings incomplete especially for insects. Family allocations included; no authorities given for genus names, latter need to be added from other sources as available.
20/09/2006	P. Bock Bryozoa Home Page (2006 version)[46]	130 additional families	
17/10/2006	Museum Victoria KEmu database courtesy Dr. E. Wallis, Museum Victoria[47]	9,300 additional extant+fossil genus names, plus 56,300 species names	Includes both extant and fossil genus plus species names.
6/11/2006	Martin & Davis 2001 Crustacea classification (Martin and Davis 2001)	130 additional families	
21/11/2006	Print work: "The Fossil Record 2" (Benton 1993) as represented by derived spreadsheet available via the Fossil Record 2 website[48]	3,200 (of 5,600) additional, largely fossil families	Extant/fossil flags also upgraded for families already held, as needed

[45] http://sn2000.taxonomy.nl/, see also http://taxonomicon.taxonomy.nl/
[46] http://www.bryozoa.net/
[47] Online version subsequently available at http://collections.museumvictoria.com.au/
[48] http://palaeo.gly.bris.ac.uk/fossilrecord2/fossilrecord/

Date	Source	Counts	Notes
7/12/2006	Aphia database (2006 version), courtesy Dr. E. Vanden Bergh, VLIZ	119 additional families, 3,300 additional genus names and 45,000 additional species names	Includes the European Register of Marine Species (ERMS) plus a number of other regional data compilations.
6/02/2007	Sepkoski (2002) compilation of marine fossil animal genera, via a publicly available online data summary[49]	27,000 additional marine fossil animal genus names	No authorities given; genera classified to order only. Sepkoski fossil flag and geologic range added for all included genus names including the subset already held.
29/05/2007	*Index Nominum Genericorum* (2007 version), courtesy Dr. E. Farr, Smithsonian Institution (online version at[50])	401 additional families, 36,000 additional extant+fossil botanical genus names	Genus names are allocated to family in most cases, the latter subsequently adjusted using other sources e.g., APG III and GRIN.
11/06/2007	The phytoPal Taxonomic Database, Feb 2007 version (Mullins 2007)	400 additional genera, 5,500 species names	Useful source for acritarch genera and species (all fossil).
23/08/2007	Australian Faunal Directory (2007 version)[51]	192 additional families, 9,800 additional genus names, 55,000 additional species names	
20/09/2007	Species 2000 New Zealand text file (pre-publication version), courtesy Dr. D. Gordon, NIWA, New Zealand (subsequently published as Gordon 2009-2011)	54 additional families, 1,800 additional genus names, 10,600 additional species names	
24/06/2008	Euzéby, List of Prokaryotic names with Standing in Nomenclature (LPSN)[52], 2008 version	77 additional families, 450 additional genus names	Species names not uploaded (many/most will be in CoL 2006).
14/04/2009	*Nomenclator Zoologicus*, digital version as compiled by the uBio project[53]	394 additional families, 205,000 additional animal genus/subgenus names	Names have authorities and microcitations, no families. Cited authors, publication information also used to upgrade c. 120,000 genus names already held. Data were pre-processed prior to upload to introduce corrections as noted, separate *nomina nuda* from cited first instance of valid publication, and remove some OCR errors.

[49] http://strata.geology.wisc.edu/jack/
[50] http://botany.si.edu/ing/
[51] http://www.environment.gov.au/biodiversity/abrs/online-resources/fauna/afd/
[52] http://www.bacterio.net/
[53] http://uio.mbl.edu/NomenclatorZoologicus/

Date	Source	Count	Notes
12/05/2009	Eschmeyer Catalog of Fishes[54], online 2009 version	800 additional genus names	Family allocations for names already held updated as needed, genus synonym status also checked. Species names not uploaded (many/most will be in CoL 2006).
13/07/2009	*Index Fungorum*[55], online 2009 version	171 additional families, 1,800 additional genus names	Family allocations for names already held updated as needed, genus synonym status also checked. Species names not uploaded (many/most will be in CoL 2006).
17/02/2010	Classification of living and fossil genera of decapod crustaceans (De Grave et al. 2009)	75 additional families, 330 additional genus names	Family allocations for names already held updated as needed.
14/04/2010	GBIF data—various including PaleoDB, Yale University Peabody Museum, others; data courtesy D. Remsen, GBIF.	1,100 additional families	
22/04/2010	Classification and nomenclator of gastropod families (Bouchet and Rocroi 2005)	92 additional families	
6/12/2010	APG III classification (Angiosperm Phylogeny Group 2009)	39 additional families	Selected angiosperm families merged as per this treatment.
19/01/2011	The Plant List, version 1.0[56]	1,400 additional genus names	New genus names checked against other sources (IPNI, Tropicos) for authorities (not included in The Plant List) and associated microcitations. Species names not uploaded.
29/04/2011	Animal genus names and original works as reported in search of "Zoological Record"[57] for newly published names, 2005-2011, plus supplementary web checks for missing diacritics and correct authority citation	4,000 additional genus names	Bibliographic details for articles also uploaded, to accompany new genus names.
9/05/2011	New fossil plant genera as published in works listed in the Kansas University online Bibliography of Paleobotany[58]	463 additional genus names	Bibliographic details for articles also uploaded, to accompany new genus names.

[54] http://www.calacademy.org/scientists/projects/catalog-of-fishes
[55] http://www.indexfungorum.org/
[56] Originally at http://www.theplantlist.org/, archived copy at http://www.theplantlist.org/1/.
[57] http://wokinfo.com/products_tools/specialized/zr/
[58] http://paleobotany.bio.ku.edu/BiblioOfPaleo.htm

Date	Source	Additional names	Notes
3/8/2011	ICTV Virus database (2011 edition), downloaded from[59]	90 additional genus names	
7/10/2011	Germplasm Resources Information Network (GRIN) Taxonomy for Plants (Sep 2011 version)[60]	1,600 additional genus names	All extant higher plant genera updated to follow GRIN family allocations as well as current name/synonym status.
4/05/2012	Hallan Biology Catalog (online version as at Mar 2012)[61]	728 additional families, 9,400 additional genus names, 208,000 additional species names	
26/06/2012	Eschmeyer online (Jun 2012 version)	650 additional genus names	Update only (for names not in 2009 version), genera only.
29/8/2012	Index Nominum Genericorum (Jul 2012 version)	700 additional genus names	Update only (for names not in 2007 version).
9/04/2013	Index Fungorum (update supplied courtesy P. Kirk)	600 additional genus names	Update only (for names not in 2009 version), genera only.
3/06/2013	WoRMS[62] (Mar 2013 version)	1,150 additional families, 3,400 additional genera, and 226,000 additional species names	
6/05/2014	The Plant List v.1.1[63]	10 additional families, 100 additional genera	Species names not uploaded.
8/06/2015	ION data dump via BioNames (data to 2012)[64], plus supplementary web checks as required	9,600 additional genus names	Parsed for genus and/or subgenus names not previously held, date range 1990-2012. Titles of original publications also uploaded, to accompany relevant new genus name records.
6/01/2016	ION data dump via BioNames (data to 2014)[65], plus supplementary web checks as required	5,700 additional genus names	Update only (for new names since 2012 version), genera only.
13/09/2016	Review of a subset of genera presently lacking an extant/fossil flag	Extant/fossil status added for 16,000 genus names previously lacking this information	Around 1,000 names researched individually, others inferred from author/publication details.

[59] http://www.ictvonline.org/ (as at 2011)
[60] Data downloaded from ftp://ftp.ars-grin.gov/pub/misc/tax, apparently no longer available in that form. Current version searchable online via https://mpgsweb.ars-grin.gov/gringlobal/taxon/taxonomysearch.aspx.
[61] http://bug.tamu.edu/research/collection/hallan/
[62] http://www.marinespecies.org/
[63] http://www.theplantlist.org/ new version September 2013
[64] Downloaded from http://www.bionames.org/data in August 2014—no longer available as a discrete download, but content is included in 2015 version below, with additional data
[65] http://www.bionames.org/data (April 2015 version), also available at https://zenodo.org/record/16863. Original data sourced from ION, http://www.organismnames.com.

4/10/2016	WoRMS (September 2016 version)	2,300 additional genus names, plus attributes / taxonomic placement upgraded for a further 43,000 taxa.	Update only (for new names since 2013 version), genera only.
16/12/2016	APG IV (Angiosperm Phylogeny Group 2016)	5 new angiosperm orders, c. 10 new families; relevant moves, splits and mergers carried out at family level, some genera relocated.	Supersedes previous treatment from GRIN and APGIII, where different.

APPENDIX 2: IRMNG DATA CONVENTIONS (2006-CURRENT)

Scientific names

- Scientific names incorporating diacritical marks (other than the optional diaeresis in botany) are not permitted under the relevant nomenclatural Codes, refer International Commission on Zoological Nomenclature (1999) Article 27 (zoology); McNeill et al. (2012) Article 60.6 (botany). Where supplied (sometimes via older works) they are standardized according to rules specified in the botanical Code (ICNafp), such that *mülleri* becomes *muelleri*, and so on.
  - Exception: for botanical names incorporating the diaeresis (example: *Isoëtes*, the quillwort), a "plain form" of the name (*Isoetes* in this case) is created for use within IRMNG and the original version maintained as an alternative spelling with its own entry, and listed as a synonym of the plain form / present accepted name, as applicable.
- Specific epithets with an initial capital letter (archaic / incorrect usage) are normalized to lowercase.
  - Example: *Hippopotamus Madagascariensis* is normalized to *Hippopotamus madagascariensis*.
- Subgenera are removed where supplied in parentheses between a genus name and specific epithet (refer example cited earlier, under "Treatment of subgenus names")—species are presently attached to their parent name at generic, not subgeneric rank.

Authorities

- Author surnames are preferably spelled out in full—including authors of botanical genera (not yet implemented for botanical species, but may be in the future)—and include publication year when known.
  - Examples (generic level):
    - *Homo* Linnaeus, 1758 (zoology)
    - *Prunus* Linnaeus, 1753, not "*Prunus* L." (botany).
- Zoological authors are given without initials (except as noted below), for botanical and bacteriological authors these are included when known.
  - example:
    - *Uharella* Taylor, Casadío & Gordon, 2008 (zoology);
    - *Scutifolium* D.W. Taylor, G.J. Berner & S.H. Basha, 2008 (botany).

  In zoology, an exception is made for a small number of cases where confusion is possible due to authors sharing the same surname working on the same group at a similar time, such as H. Adams vs. A. Adams (molluscs), M. Sars vs. G.O. Sars (marine crustaceans).

- Author combinations are conjoined using the ampersand character consistently.
  - Example: "Cavalier-Smith & Chao", not "Cavalier-Smith et Chao" (convention elsewhere in botanical works), "Cavalier-Smith and Chao", etc.
- "Long form" citations (i.e., author *A* in [work authorship] *A & B*, etc.) are preferred to "short form" citations, when known to be applicable.
  - Example: "Cavalier-Smith in Cavalier-Smith & Chao, 2010", not "Cavalier-Smith, 2010".
- Diacritics in author names are added back where missing in supplied data.
  - Examples: Chujo, 1969 becomes Chûjô, 1969, etc.

- Representation of non-English author names is checked against additional sources and adjusted/corrected as required.
  - Example: "Lu Junchang, Pu Hanyong, Xu Li, Wu Yanhua & Wei Xuefang, 2012" (as given in ION) is corrected to "Lu, Pu, Xu, Wu & Wei, 2012"; "Vilela Cruz, Falcao Salles & Hamada, 2013" (in ION) is corrected to "Cruz, Salles & Hamada, 2013".
- Multiple authorships are spelled out in full where these do not exceed five authors.
  - Example: "Cruz, Salles & Hamada, 2013", not "Cruz et al., 2013" ("et al." is used from the sixth author onwards).
- A comma is inserted before all dates in authorship citations (this is optional according the zoological Code but standardized for IRMNG use), e.g., "Linnaeus, 1758", not "Linnaeus 1758".
- "Curly" apostrophes are replaced with "plain" forms within author names for consistency of data entry and text searching (e.g., "O'Donohue", not "O'Donohue").

Literature citations

From 2016, literature citations are being entered according to WoRMS/Aphia conventions, with separate, searchable fields for author name(s), year, article title, journal title, volume, article pagination, and page in work, also DOI (digital object identifier) and online link as available. These are then reassembled and formatted as required (e.g., with the journal title italicized, and DOI / online link presented as hyperlinks) for presentation on relevant source and taxon pages. Citations entered prior to 2016 are slowly being converted to the new standard, commencing with those used for large numbers of taxon names or as the source for the most recent treatments for certain groups.