

Accepted for publication in *Trends in Ecology and Evolution*, 22<sup>nd</sup> August 2018

Article type: *Scientific Life*

## **Sustainable biodiversity databasing: international, collaborative, dynamic, centralised**

Mark J. Costello<sup>1</sup>, Tammy Horton<sup>2</sup>, Andreas Kroh<sup>3</sup>

<sup>1</sup> Institute of Marine Science, University of Auckland, Auckland 1142, New Zealand.

<sup>2</sup> National Oceanography Centre, Southampton, European Way, Southampton SO14 3ZH, UK. [tammy.horton@noc.ac.uk](mailto:tammy.horton@noc.ac.uk)

<sup>3</sup> Natural History Museum Vienna, Department of Geology and Palaeontology, Burgring 7, 1010 Vienna, Austria. [andreas.kroh@nhm-wien.ac.at](mailto:andreas.kroh@nhm-wien.ac.at)

Correspondence [m.costello@auckland.ac.nz](mailto:m.costello@auckland.ac.nz)

### **Abstract**

The World Register of Marine Species is a sustainable model of international collaboration around a centralised database that provides expert validated biodiversity data freely online. This model could be replicated for the over 1.2 million terrestrial and freshwater species to improve quality control and data management in biology and ecology globally.

**Keywords:** taxonomy, nomenclature, species, informatics, classification, biodiversity, standardisation, databases.

## **Biodiversity Informatics**

A natural consequence of the age of information technology has been the emergence of biodiversity informatics and its associated software tools and data systems that transform the speed and capacity to do research [1]. However, such tools are entirely dependent on the quality and quantity of the resource's content. In biology, the key biological standard is the species concept, and the primary data are where and when species have occurred.

One of the greatest problems in biological and ecological sciences is that there are at least 1.5 million named species, but several times more scientific names than biological species, plus numerous misspellings and conflicting classifications that compromise data management and complicate usage of species data by non-taxonomists [2]. For example, only one third of over one million scientific names of plants are accepted for use (<http://www.theplantlist.org>). These problems are amplified when we try to compare data across studies, geographically and over time. Society would benefit from increased effort in biodiversity science to describe all species and quality assure taxonomy, but this needs collaboration and coordination [2, 3]. The cost of this ignorance due to loss of biodiversity and food production (e.g. over-fishing), and mistakes in natural resource management, is likely to be in US\$ billions per year [3]. Here we use WoRMS as an exemplar to show that collaboration around a centralised dynamic taxonomic database can significantly advance this effort.

## **Changing nomenclature and classification**

In addition to correcting errors in the literature, species names and their classification change because new knowledge results in new species, taxonomic revisions that discover synonyms, and re-classification of existing species. Users of taxonomic data find it challenging to keep track of such changes, especially when conflicting information is found in published literature

and online, and it can be difficult to access some publications. This can create problems in conservation where a 'species' may receive greater or lesser protection depending on its taxonomic status [4], and in the management of databases that use species names.

Important online nomenclatures capture the names of species, such as the Index of Organism Names (mostly animal names and includes the former Zoological Record), The Plant List, Index Fungorum, MycoBank, and ZooBank. However, with the exception of The Plant List, these are not edited by experts to be comprehensive and clarify which names are accepted, and none are continuously updated.

Users of species names include individuals, institutes, and other globally important scientific databases, such as GenBank (<https://www.ncbi.nlm.nih.gov/genbank>), Barcode of Life Database (<http://www.boldsystems.org>), [Map of Life](#), [Tree of Life](#), traits databases (e.g. [Try](#)), [FishBase](#), the [Global Biodiversity Information Facility](#), and [Ocean Biogeographic Information System](#). It is essential for such users to know which names are valid, which invalid and which are questionable, and how the application of a name has changed over time [5]. Indeed, making quality-assured expert knowledge easily accessible is one of the challenges facing science and society in many fields. These kinds of biodiversity databases depend on a dynamic, updated expert-validated taxonomy of species names. However, they are not designed technically and socially to provide such a service. GBIF and OBIS operate distributed models that publish datasets requiring checking of taxonomic names, but to date have been conveyors of such content rather than validating or curating it themselves [6].

The present system of ad hoc publications of taxonomic information and distributed online resources cannot efficiently address this challenge. Publications are often behind a paywall

rather than open access and are not easily searched. Moreover, individual papers typically deal with a subset of species of a taxon, making it necessary to extract information from multiple publications that may employ conflicting classifications and different data standards. Online name catalogues produced and maintained by one or a few experts have a high risk of stalling when their champions become unable to contribute further due to lack of funding, illness, or other factors. Succession planning may start too late and be compromised by the idiosyncratic structure of the database and need for financial support to maintain it. Owners of such resources may lack the professional informatics expertise to manage and future-proof such databases. However, one model is overcoming these challenges within the taxonomic community, particularly for those working on marine species and their relatives.

### **Benefits of a dynamic collaborative database**

The “World Register of Marine Species” (WoRMS) has now been established for over ten years [7] (Box 1). WoRMS currently contains over 243,000 valid marine species and 600,000 marine taxonomic names, and has over 130,000 unique users per month. It is updated continuously, with amendments typically occurring every few minutes, and archived monthly with a DOI (Digital Object Identifier). Despite its origins, it now also includes freshwater and terrestrial relatives of some taxa, including Mollusca, Porifera, Polychaeta and Crustacea. Its content is gradually expanding to include geographic information, images, introduced and conservation status, fossils and species traits [8]. WoRMS has been meticulous in documentation of data sources. Any information item entered in WoRMS can be linked to one or more sources, providing for traceability of decisions and offering another layer for data users to consult.

WoRMS and associated databases content are managed by almost 500 invited experts, about 300 of whom are taxonomists [7, 9, 14]. The ingredients for its sustainability include this social network of experts and a professional data management team, and a clear agreement that intellectual property rights of the content rest with the Editorial Board and are managed by its Steering Committee (Box 1) [10]. Involvement of IT-experts allows scientists to focus on their expertise and enables them to reach out for help when needed. Our experience has been that to engage experts from different countries and communities requires clarity on resource ownership, management, succession planning, and peer to peer engagement (i.e. community building). Experts are more likely to spend their time on a database if it is open access, easy to use, sustainable, quality assured, and they get most recognition when it is considered meritorious and prestigious within their community (Table 1). They, and in some case their employers, are also more willing to provide their time as editors if the product directly contributes to society, that other institutions or individuals do not profit on their work (e.g., research funding, authorships), and they have a legal right to influence its management (e.g. electing a Steering Committee). Thus, a global scale, community owned and governed, open access, continuously edited, centralised database, with professional informatics support, can provide a cost-effective sustainable service to science and society (Table 1). Other community driven bottom-up efforts, like Wikispecies and its associated Wikidata also deserve support. The same is true for initiatives driven by the scientific community like the Tree of Life web project, which is an excellent expert edited introduction to biodiversity and valuable tool for teaching, but comes in a narrative format that precludes automated analysis.

Although the conventional process of peer-reviewed literature serves the taxonomic community well, access to an expert-edited database where mistakes and omissions can be promptly corrected, and new knowledge made freely available is widely appreciated by the

broader scientific community, as evident from the over 5,000 citations of WoRMS in Google Scholar. Other online biodiversity resources, including Species 2000's Catalogue of Life, Wikispecies, Encyclopedia of Life, Global Biodiversity Information Facility and Ocean Biogeographic Information System, use information on marine species from WoRMS to support their information systems; demonstrating good cooperation within the biodiversity informatics community. Although the first two of these are expert edited, CoL lacks a centralised database and with the exception of data provided by WoRMS, most of its datasets are irregularly updated. However, the current GBIF [implementation plan](#) involves collaboration with CoL and others to provide a complete, literature referenced, automated and expert validated world species list (D. Hobern, pers. comm.).

We suggest the WoRMS model of international collaboration should be used to provide a quality-assured taxonomy for all species on Earth, and support other biodiversity related databases to make expert knowledge openly available to society. Additional benefits of this collaboration are that it becomes easier to conduct global syntheses of taxa because the information is standardised in the database [8]. For example, WoRMS enabled a world synthesis of how many marine species are named and might exist, e.g., [11] (Table 1). A global scale, expert driven, collaborative and centralised open access database, could thus be available for all species on Earth, and recently called for by conservation biologists and taxonomists [4, 12]. This is essential to provide a current taxonomy for all other biodiversity databases and publications. Following this, perhaps the next gap to be filled will be an identification guide to all life on Earth that links databases with literature and images; a "key to all life".

## Box 1. Evolution of WoRMS (World Register of Marine Species).

---

The origins of WoRMS were the European Register of Marine Species (ERMS), a conceptually similar expert validated checklist of all European marine species. This was freely available online and conventionally published [13]. To avoid complications concerning intellectual property rights with so many editors and authors, the ownership of ERMS was vested in a non-profit limited liability legally incorporated scientific society; the Society for the Management of Electronic Biodiversity Data Ltd (SMEBD). SMEBD had a deliberately broad remit to support other electronic biodiversity databases (at least in Europe). However, as these did not transpire its responsibilities were later transferred to the Editorial Board of WoRMS, which includes editors of ERMS and an increasing number of child databases with regional, taxon specific or thematic (e.g. introduced species) foci. Instead of supporting core operations, financial contributions from research grants and donations from users are used to fill gaps in and expand WoRMS content.

Shortly after the completion of ERMS, the Flanders Marine Institute (VLIZ) Data Centre offered to host it because having validated species names was critical to its marine biological data management. This relieved the editors of needing to worry about the informatics aspects of a biodiversity database. The hardware, software and 24-7 online access were now permanently supported by a team of data management specialists at a professional, government-funded data centre. This security encouraged more experts to become editors and provided the computer platform for WoRMS [14]. The involvement of a core group of taxonomists in the Census of Marine Life and its Ocean Biogeographic Information System provided the social network to expand ERMS to WoRMS [15]. WoRMS is now governed by a committee elected by its editors which has a formal agreement with the database host institute [7].

---



Table 1. Fifteen benefits of collaborative management of a biodiversity database.

---

**To science**

1. Improved quality control in biodiversity science.
2. Gaps in knowledge are more visible and encourage researchers and funding agencies to fill them.
3. Reduction in misspellings and incorrect use of species nomenclature.
4. Rapid conversion of outdated species names into state-of-the-art name, particularly outside a researchers own field of expertise.
5. Easy access to the reference for the original species description, many of which are rarely cited elsewhere.
6. Standardisation and integration makes it easier to conduct global syntheses of information.

**To users**

1. Ease of access to an electronic, standardised, authoritative species list that is classified hierarchically for use in their own data management.
2. Automated tools to classify and check spelling of species names.
3. More time-efficient to consult a single authoritative source than to research and assess accuracy of numerous disparate sources.
4. Contact details of experts easily found.
5. Relieves initiatives focussing on other (non-taxonomic) aspects of biodiversity of the need to keep track of taxonomic changes themselves.

**To the database editors**

1. Focused collaboration with colleagues internationally that can aid their personal knowledge and know-how.
  2. Information organised and archived reducing the need for the expert to have his or her own database.
  3. Citable electronic publication that is digitally archived.
  4. Peer-recognition; it is prestigious to be invited as an expert to edit the database.
-

## References

---

- 1 Hobern D, *et al.* 2014. Advancing online databases and information systems for biodiversity conservation. *Biological Conservation* 173, 65–67.
- 2 Costello MJ, *et al.* 2013. Can we name Earth’s species before they go extinct? *Science* 339, 413-416.
- 3 Wheeler, Q.D., *et al.* 2012. Mapping the biosphere: exploring species to understand the origin, organization and sustainability of biodiversity. *Systematics and biodiversity* 10 (1), 1-20.
- 4 Garnett S.T., Christidis L. 2017. Taxonomy anarchy hampers conservation. *Nature* 546: 25–27.
- 5 Costello M.J., *et al.* 2013. Global coordination and standardisation in marine biodiversity through the World Register of Marine Species (WoRMS) and related databases. *PLoS ONE* 8(1): e51629.
- 6 Costello MJ, *et al.* 2013. Data should be published, cited and peer-reviewed. *Trends in Ecology and Evolution* 28 (8), 454-461.
- 7 Vandepitte L, *et al.* (2018) A decade of the World Register of Marine Species - General insights and experiences from the Data Management Team: Where are we, what have we learned and how can we continue? *PLoS ONE* 13(4): e0194599.  
<https://doi.org/10.1371/journal.pone.0194599>
- 8 Costello M.J., *et al.* 2015. Biological and ecological traits of marine species. *PeerJ*, e1201.
- 9 Horton T. *et al.* 2017. Improving nomenclatural consistency: a decade of experience in the World Register of Marine Species. *European Journal of Taxonomy*, No. 389. DOI: <https://doi.org/10.5852/ejt.2017.389>.
- 10 Costello M.J., *et al.* 2014. Strategies for the sustainability of online open-access biodiversity databases. *Biological Conservation* 173, 155-165.
- 11 Appeltans W, *et al.* 2012. The magnitude of global marine species diversity. *Current Biology* 22, 1-14.
- 12 Thomson, S.A., *et al.* 2018. Taxonomy based on science is necessary for global conservation. *PLoS Biology* 16(3), p.e2005075.
- 13 Costello, M. J., *et al.* (editors) 2001. European Register of Marine Species. A check-list of marine species in Europe and a bibliography of guides to their identification. *Patrimoines naturels* 50, 1-463.
- 14 Vandepitte L., *et al.* 2015. How Aphia—the platform behind several online and taxonomically oriented databases—can serve both the taxonomic community and the field of biodiversity informatics. *Journal of Marine Science Engineering* 3: 1448-1473.
- 15 O’Dor R., *et al.* 2012. A Census of Fishes – and everything they eat: How the Census of Marine Life advanced fisheries science. *Fisheries* 37:9, 398-409.