Opinion

Biodiversity data should be published, cited, and peer reviewed

Mark J. Costello[1], William K. Michener[2], Mark Gahegan[3], Zhi-Qiang Zhang[4], and Philip E. Bourne[5]

[1] Institute of Marine Science, University of Auckland, Auckland, 1142, New Zealand
[2] University Libraries, University of New Mexico, Albuquerque, NM 87131-0001, USA
[3] Centre for eResearch, University of Auckland, Auckland, 1142, New Zealand
[4] Landcare Research, 231 Morrin Road, Auckland, 1072, New Zealand
[5] Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, La Jolla, CA, 92093-0657, USA

Concerns over data quality impede the use of public biodiversity databases and subsequent benefits to society. Data publication could follow the well-established publication process: with automated quality checks, peer review, and editorial decisions. This would improve data accuracy, reduce the need for users to 'clean' the data, and might increase data use. Authors and editors would get due credit for a peer-reviewed (data) publication through use and citation metrics. Adopting standards related to data citation, accessibility, metadata, and quality control would facilitate integration of data across data sets. Here, we propose a staged publication process involving editorial and technical quality controls, of which the final (and optional) stage includes peer review, the most meritorious publication standard in science.

The importance of biodiversity data

Biodiversity data

In today's digital world, all biodiversity information and data should be available online, unless there are sound reasons why they should be kept confidential (e.g., nesting site of a rare bird). Information that is not online will be overlooked. For biodiversity data, the requisite storage capacity and infrastructure are available, and there are continuing improvements in data management tools [1,2]. However, quality assurance is inconsistent and a culture of data publication is lacking. Consequently, few scientists use biodiversity databases for their research, and fewer still contribute data back to the community. Meanwhile, publicly funded data are 'lost', and global issues that threaten human food sources and ecosystem health remain, such as climate change, overfishing, infectious diseases, and invasive species. Addressing these challenges requires that existing data be properly maintained, trusted, and unconditionally accessible [3,4].

Biodiversity data can include inventories of species names and synonyms, species distributions, images and sounds, ecological interactions, behaviour, data set descriptions, and analyses and interpretations [5]. Here, we are most concerned with the primary biodiversity data rather than the secondary (e.g., modelled or simulated) data derived from them, and interpretations and descriptions around data. Thus, data can be numerical, categorical (e.g., species or place names), images, or sounds.

The rate at which new data are published through the Global Biodiversity Information Facility (GBIF) (Box 1), as a proportion of available data, is declining each year [6]. GBIF was established to make biodiversity data publicly available and, thus, to satisfy a key aim of the Convention on Biological Diversity. Nonetheless, more data are continually being collected [7–9]. Moreover, centuries of irreplaceable historic data on biodiversity and the environment need to be digitised to provide the historical context for present observations, and enable predictive modelling of the consequences of human activities for the environment and biodiversity [10–13]. This historic record is especially important for taxonomy, because the first description of a species has legal priority for the name of that species [14,15].

Motivating data publication

It is necessary to motivate and reward the contribution of data to international integrated databases by bringing such data into the mainstream of respected scientific publication [5,9,16,17]. Data publication increases the visibility of scientists' work and citation rates [18]. This can be an incentive to some scientists, but still less than half of authors make their data publicly available online [18,19]. The situation in ecology may be worse; a survey of environmental biology publications from 2005 to 2009 found that 57% had not released their data and, when genetic data were excluded, only 8% had [20]. Even in those journals that require that data be made available, one study found that most (59%) papers did not submit their data [21]. Most scientists (92%) agree that data sharing is important [22]. Smit [23] found that, whereas 80% of scientists wanted access to data created by others, 13% did not want to share their data and only 20% have actually shared data. Clearly, data-sharing agreements and policies are insufficient, and new approaches are required [5].

Data publication

Decades ago, journals frequently published species inventories, ecological survey data, and data appendices.

Corresponding author: Costello, M.J. (m.costello@auckland.ac.nz).
Keywords: databases; species; journals; quality control; Global Biodiversity Information Facility.

0169-5347/$ – see front matter
 http://dx.doi.org/10.1016/j.tree.2013.05.002

Box 1. Biodiversity data publication by the Global Biodiversity Information Facility

Over its first decade, GBIF published over 370 million records of species, from 12 000 data sets supplied by 400 organisations from over 40 countries, with over 4.5 million names (Figure I). The names include scientific, vernacular, and other names, and amount to almost 1 million species, of which 590 000 have distribution data (Tim Robertson, personal communication). The marine component of GBIF, OBIS, contains over 120 000 species, which is over half of all described marine species [61–63]. Approximately 80% of records represent species observations and samples rather than museum specimens [9]. The data from each source are integrated into a large searchable database [53]. Over 85% of animals and 76% of plant species can be mapped [6]. Thus, the sum of local and regional data can be used to examine global-scale phenomena. Over two-thirds of the data sets in GBIF have been provided by government organisations whose staff are directed to do so. Far fewer data sets are delivered from the academic community, although it publishes approximately 75% of all scientific papers, despite comprising only 15–50% of all scientists [38]. Nevertheless, the number of publications that has used data from GBIF is increasing (Figure I).

GBIF needs to address not only the amount of data, but also the geographic, temporal, and taxonomic coverage, and accuracy (quality). Scientists' concerns over data accuracy might be impeding data reuse and consequent benefits to society [22,64]. A more incentivised publication model could encourage scientists to offer data sets to GBIF for publication, just as they now offer papers to journals to publish. This could be direct to GBIF, through one of the GBIF participants, or offered through a biodiversity journal. This does not exclude the present process of data publication continuing, but offers a quality-assured process that might be more attractive to some scientists and data users.

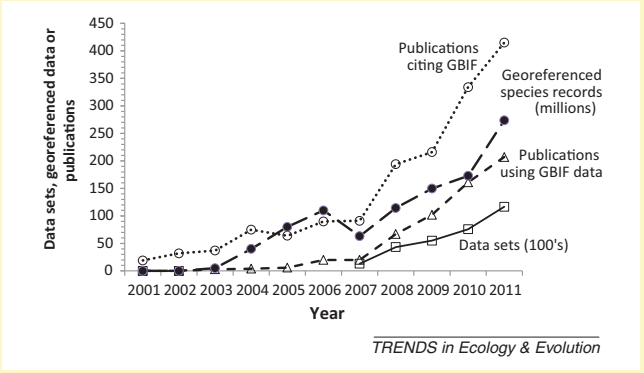

Figure I. The increasing number of millions of species distribution records published by the Global Biodiversity Information Facility (GBIF) (solid circles), hundreds of data sets (open squares), publications that use data from GBIF (open triangles), and publications that cite GBIF (open circles). Data from GBIF [65].

However, printing and postage costs led to journals being reluctant to publish tables and appendices of primary data. Today, the availability of online appendices and electronic publication means that this should no longer be an issue, and some biodiversity journals (e.g., *Zootaxa* and *Phytotaxa*) publish species inventories both in print and online.

It is increasingly acknowledged that data created using public funds or for the public good (e.g., environmental monitoring) should be publicly available [5,24,25]. Likewise, many publishers expect authors to make their data publicly available, ideally in international databases, in permanent institutional repositories, or as online supplementary material (reviewed in [5]). However, peer review and editorial processes generally exclude assessment of such data. Important exceptions include *Data Papers* and *Ecological Monographs* of the Ecological Society of America, the *Earth System Science Data Journal*, *BioInvasions Records*, and *Datasets in Ecology*. Also, the publisher PenSoft has announced the introduction of 'data papers' in six of its journals [26]. However, unless authors publish in a specialist 'data journal', there is often no oversight to ensure that the data set adheres to accepted standards, has adequate metadata, and is largely error free.

Published online appendices are not ideal because they are not usually peer reviewed [27], subject to independent editorial attention, and may not be open access. Because such appendices are not required to conform to data and metadata standards, their reuse can be problematic. Furthermore, much 'supplemental material' is not permanently archived and can become inaccessible over time [28,29]. Although print publications with an ISSN and ISBN are archived in libraries, this is not the case for online supplementary material. Institutional repositories can be preferable where they provide permanent archiving, but most lack peer review, editorial review, and alignment with emerging disciplinary standards. A better option is to deposit data in Dryad (http://datadryad.org) because it is a centralised open-access repository directly linked to journals. By early 2013, it had published over 7000 data files from articles published in 187 journals. Some journals now require authors to pre-deposit data in Dryad rather than as 'supporting material' on the journal website. However, the data are not subject to independent quality checks, are not required to conform to particular standards, are not peer reviewed, and are limited to data associated with published papers. By comparison, far more biodiversity data are published through GBIF by government organisations, of which only fragments may be associated with research papers.

In contrast to journals, specialised data centres are most familiar with data standards, and in-house staff typically provide some quality assurance of data and metadata (e.g., PANGAEA, the Distributed Active Archive Centers of the National Aeronautics and Space Administration, GenBank, and Protein Data Bank). Thus, specialised data centres are preferable for data publication.

The problem with 'data sharing'

Perhaps the primary reason why data publication is not the norm is that most data policies refer to 'sharing' or making data 'available', rather than 'publishing' (e.g., [30,31]). This is a key distinction, because making data available suggests a negotiation between the parties involved as to the terms and conditions of availability. This might require direct payment, joint authorship, or partnership in research contracts (e.g., [5,24,32,33]). Fortunately, this is not the case for scientific papers, and should also not be so for data sets [34]. Calls for making data 'available' can be counter-productive because they pressure scientists to do something outside their comfort zone: for example, they may not have clarified data ownership and a dissemination policy with their collaborators,