Opinion

Cell PRESS

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.

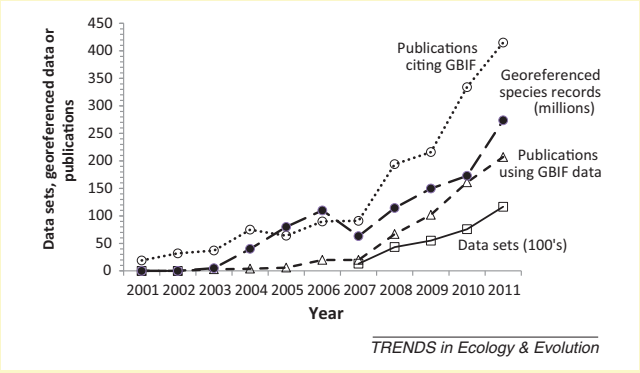

Figure I. The increasing number of millions of species distribution records published by the Global Biodiversity Information Facility (GBIF) (solid circles), hundreds of data sets (open squares), publications that use data from GBIF (open triangles), and publications that cite GBIF (open circles). Data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

Box 2. Publishing Issues

Implementing data peer review

Peer review could include a list of questions for referees, such as: is the data set description complete, clear, and adequate to understand the taxonomic, temporal, and geographic scope of the data? Does it contain appropriate descriptions and citation of methods and data analyses? Do the data adhere to community standards? How might the data be used by other researchers? How significant is the data set in terms of size, scope, and uniqueness? The publisher Pensoft has developed its own guidelines for such peer review [26].

A concern in adopting peer review is the availability of willing referees, given that this is already a problem for science publications. It is remarkable that there are few incentives used to attract referees, yet most scientists provide their time gratis. Incentives used by subscription-paid journals, such as temporary free access to publications online, are not available to open-access publications. Nevertheless, other options can be explored, such as public acknowledgement of the referees, invitations to write special articles and join editorial boards, payment of honoraria, and employment of a few experts as 'in-house' reviewers instead of relying on many unpaid referees.

Open access

'Open access' indicates that biodiversity data are freely accessible to countries, including developing countries, where the data might have been collected. It means that third parties are expected to use the data, create new data sets from it, and benefit from it in terms of their research, making policy decisions, or developing commercial applications from the data. Such uses should be seen as signs of success and justification for the funds that enabled its publication. Having collective databases such as GBIF, OBIS, and VertNet, is simpler in terms of user access, and for the development of standards and analytical tools that facilitate data integration and synthesis. Such collective resources will also be more cost effective to support, and precedents for tiered financial contributions have already been established (e.g., based on gross domestic product). Authors of open-access publications are more cited [55,66], and the open-access business model is more cost effective for society [67,68]. The standardised Creative Commons licences are often used for open-access databases and publications. Journal websites are not suitable repositories for data because they are not permanent archives, do not require standards, do not subject data to any quality checks, and most are not open access. However, somebody has to pay for open access.

Publication costs

Who pays for the publication process and long-term maintenance of access to the data? Open-access journals typically charge the authors, thereby excluding those who cannot pay. This cost might come from readers (users), institutional libraries, authors, or be sponsored by organisations such as government institutions. There are additional costs when data are integrated with other data sets, a service provided by government-funded data centres, such as GBIF. Considering the importance of biodiversity data being open access to society, it is appropriate that data publication is government funded.

employer, or funding sources. Furthermore, significant work can be required to get the data into a well-described format that others can use. Whether these concerns are justified is immaterial, because there is little incentive for the scientist to spend time overcoming them when their success is primarily judged by publications.

Biodiversity data publication

There are many open-access scholarly biodiversity databases (e.g., listed in [24]). Most provide information on species such as FishBase [35], AlgaeBase [36], and the Global Invasive Species Database (www.issg.org/database), and images, such as to identify individual whales from their photographs (www.cetabase.info). A few provide standardised distribution data, such as the Ocean Biogeographic Information System (OBIS [37]), which republishes data into GBIF. Both GBIF and OBIS have inter-governmental governance and funding structures, and are organised as a global network of nodes that publish data through a single portal.

Quality assurance and control

Data sets are often accompanied by statements about the publisher and creators not being responsible for the use others might make of the data, or for any errors contained therein. In a 2011 survey of GBIF participants, 57% of the 35 respondents stated they could not guarantee data quality and 43% had no statement about data quality. Such disclaimers are not prominent in conventional scientific journals because authors and publishers are responsible for ensuring high-quality publications. This often includes a prior review of the submission by editors and independent experts (i.e., peer review). Furthermore, if after publication, some errors, plagiarism, or other defects are found, then they can be corrected or, in extreme cases, the publication can be withdrawn. It is clear to the scientific community and the public that different publications have been subject to different levels of quality assurance and control (QA/QC), of which peer review is the highest quality mark (Box 2). Thus, instead of having a disclaimer, data publishers, similar to scientific journals, should be proactive and use transparent QA/QC procedures.

Peer review is integral to science [38]. Non-peer-reviewed publications are regarded as inferior by scientists, their employers, and policy-makers [39]. Publications that are not peer reviewed have negligible value in metrics of scientists' productivity and reputation. A review for the European Union recognised the rapidly growing data volume, but made little mention of the need to capture past data and methods of quality control, and no mention of the need for peer review [17]. Although the question of peer review of data publications has not been considered in detail, it is now being encouraged [5,25,27,40,41]. The fact that species web page profiles in the Marine Life Information Network and the Global Invasive Species Database, Global Species Databases published by Species 2000, *Data Papers* published by the Ecological Society of America, species inventories in *Zootaxa*, and data in the NASA Planetary Data System [27] are peer reviewed demonstrates that it is possible.

Fitness for purpose

Whereas in the early stages of collating data, the emphasis might be on the quantity of data, quality must be checked. For example, during the early years of the Protein Data Bank, one-third of its budget was spent on data cleaning. Considering that all potential data uses are unlikely to be predicted and that most data might be useful for some purposes, there can never be too much biodiversity data. However, establishing that data are fit for a specific purpose is often a difficult task, and can entail study of both metadata and the processes (workflow) used to create the

data, as well as data content. Enhancements to metadata should be driven by the need to help users understand appropriate uses of the data.

Data set citation

Calls for data sets to be cited in a conventional manner are now widespread (e.g., [5,16,27,32–34,42–46]), and an online register that links data sets to digital object identifiers (DOI) has been launched by DataCite (http://datacite.org). However, few data sets and online resources have been cited in this way [39]. Nearly all scientists want credit for the use of their data [23]. In the process proposed here, editors determine the citation style for their journal, but one can expect it to include the common elements of authors, title, publisher, and publication date [5,34,39]. Costello [5] listed 16 benefits of data publication, but nine of these can only be realised if the data are cited in this way. Following an established publication process implies standard citation of data sets, citation tracking, permanent archiving, and other use metrics [1,5,16].

Journals and authors presently have different policies on how to cite online resources. Some include a universal resource locator (URL) in the paper text, rather than citation in the bibliography or references. Furthermore, the practice of citing the date on which an online resource was accessed is only appropriate when it is a web page that can change over time [34]. The conventional publication of data sets, paralleling the 'papers' in a journal, would make it clear that they should be cited in the references. Thus, OBIS proposed a citation as part of the metadata for the data sets it published in 2006 [47] and various options have been considered by GBIF [48].

Citations should not be confused with codes for tracking publications, data, or parts of publications; but these can be added to citations. Such codes include DOI, Life Science Identifiers, and Uniform Resource Names that aid databases in tracking publication citations (e.g., [49]) and the provenance of individual data items [14,50]. This provides new opportunities to develop data-use metrics that measure the impact of data publication. Various data-use metrics might be necessary because data sets might not always be cited and tracked by scientific abstracting services [1].

When a data paper links to the source of a data set or database, typically a URL is included. However, a URL can

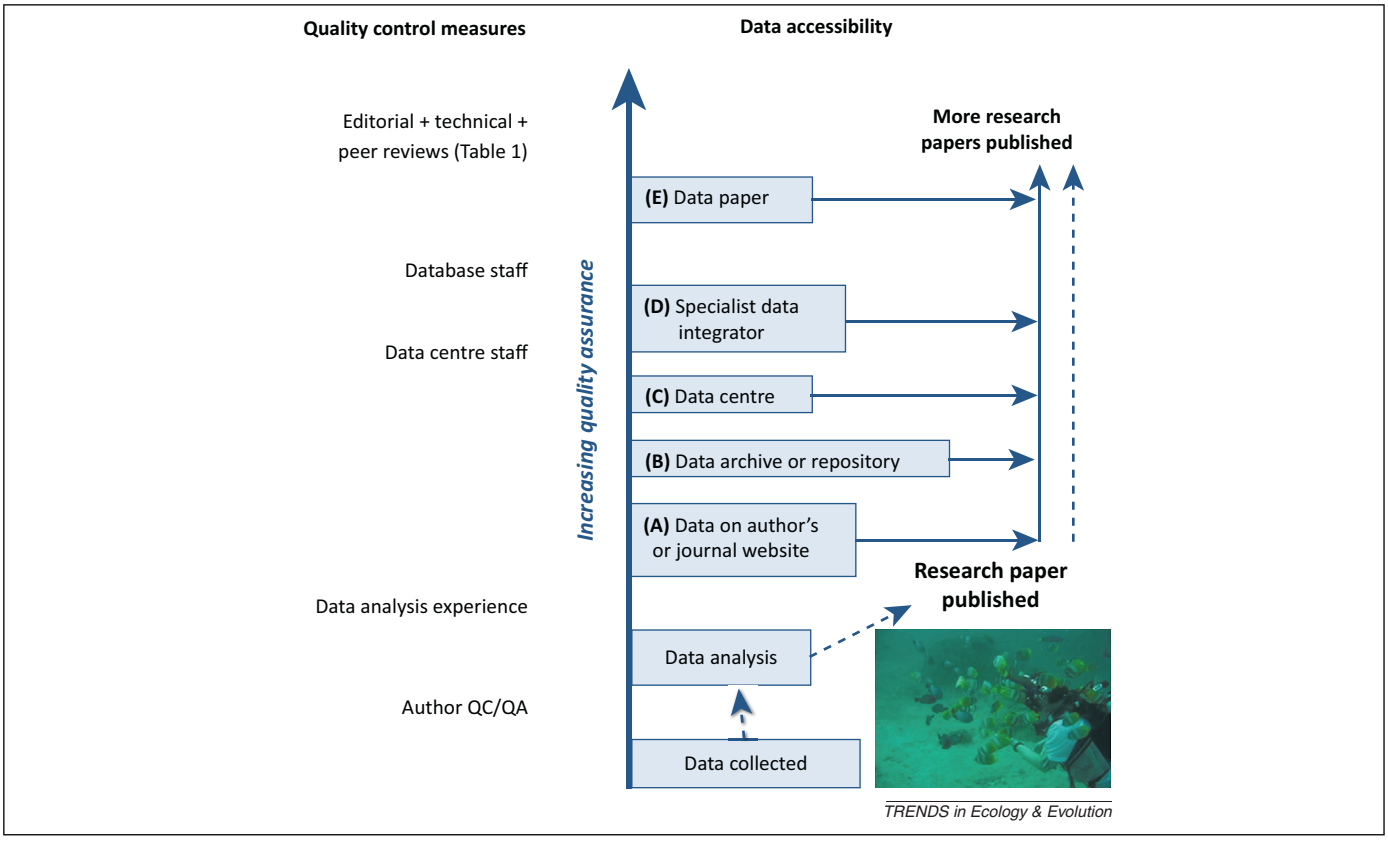

Figure 1. The workflow involved in biodiversity data publication has multiple opportunities for improving data and metadata quality (left panel) and accessibility (right panel). First, the conventional research publication process from data collection, to analysis, and publication of the interpretation of the data in a paper is illustrated with broken lines. Associated data may be published online (solid lines) through **(A)** the author's website or that of the journal; **(B)** an archive or repository without any independent quality checks or standards; **(C)** a data centre with staff who check that the data conform to particular standards; and/or **(D)** a specialist data centre that integrates the data with similar data sets. A third option, is to publish the data as a 'data paper' **(E)** that describes the data but may not analyse them beyond some descriptive statistics. In this example, data on the abundance of fish observed by scuba divers is collected, including sampling locations and dates, species counts, body size, and habitat. The data collectors would typically do some quality control (QC) on their data. During data analysis, any inconsistencies in data recording between observers are corrected, providing further quality assurance (QA). Following publication of the first paper from these analyses, the supporting data are made available on the website of the journal, the authors' website, and/or a digital archive (e.g., Dryad). These data are typically not independently reviewed or required to conform to specific standards. The authors decide to submit the data to a data centre (e.g., PANGAEA). The data centre staff inspect the data and ask the author to correct some inconsistencies, provide additional metadata, and reorganise them to conform to domain standards. Data on species distribution are published in the Global Biodiversity Information Facility (GBIF), where they are integrated with millions of other data values and can now be searched and reanalysed by others. Alternatively, the author could have submitted the data set for publication as a 'data paper', whereby the journal would have performed the full suite of technical, editorial, and peer-review checks to ensure the data and metadata are of the highest quality and maximised and may be easily integrated with GBIF and other specialist databases.

change when data sets are moved or domain names are changed. A central URL registry for published data sets (similar to what CrossRef is doing for papers) is needed so that when the URL for data sets change, the records in the registry are updated. Persistent DOIs are already being assigned to data sets (e.g., by DataCite and Dryad). GBIF and other data publishers should lead in supporting a central registry for resolving various digital identifiers for permanent linking to biodiversity data sets.

A proposed data publication process

Data publication should follow established practices of other scientific publications in several regards, including editorial quality control, independent peer-review, published data set citation, and permanent archiving (Figure 1). Metadata descriptors must include authors, their contact details, abstract, keywords, and other information necessary to enable abstracting services to include the publications in their databases [1,34]. This more formal approach to data publication will require more comprehensive metadata that are essential for appropriate data use [9], such as data provenance, context, precision, and references to papers that used the data. It is likely to require the use of metadata standards and standardised vocabularies. Michener *et al.* [9] provided a list of metadata descriptors for ecology data to which taxonomic metadata (e.g., taxa included in the study) could be easily added. Data provenance might be traceable through the executable workflow of how the data were produced or assembled. Metadata might ultimately become a domain ontology, with rich semantics describing the various categories used, and their inter-relations.

The data publication process proposed here may warrant establishment of online open-access biodiversity data journals, and the adoption by existing journals of data publication. We emphasise the importance of open access for biodiversity data and recognise this requires financial support (Box 2). This is also the principle behind many government and intergovernmental policies for data availability (reviewed in [5]), including the public funding of GBIF, GenBank, and other databases.

Data publishers

There could be several data publication journals. Some might specialise in photographs to aid species identification, species geographic distribution, or population abundance time-series data, and to develop new data standards (Box 3). They could be linked through portals with other databases, such as GBIF, and to published literature, similar to the Biodiversity Hubs created by the Public Library of Science. *ZooKeys* and *PhytoKeys* already submit data to GBIF post-publication of their accompanying paper [51]. One possible way forward, whereby a data journal could be widely abstracted and get an Impact Factor, is to design it as a peer-reviewed scholarly journal covering more than only 'data' papers, with various publications of scholarly value so that these will be abstracted and included in journal citation rankings. These might include editorials on topics of interest in the field, best practice, case studies, invited review papers on current topics in biodiversity data and databases, papers on methods and data standards, introductions to, and reviews of, new software for data exploration, presentation or analysis, and other related topics. Data paper authors could receive automated messages when their paper is downloaded and cited, and be contactable by data users so that new collaborative publications might arise. The ideal solution might be for data publication to occur through data journals that archive data in integrated databases managed by data centres, such as GBIF.

Box 3. The need for biodiversity data standards

A benefit of integrating data into one (possibly virtual) system is that it drives standards for data management. Standardised, quality assured, permanently archived databases are essential to manage the collection, storage, and accessibility of this growing data stream. This is widely recognised [17] with significant investments in new data infrastructures, but with insufficient attention to bringing past data into a quality-controlled digital environment. For example, an expert-validated inventory of all species that reconciles synonyms and nomenclatural confusion is essential for integrating high-quality biodiversity data from different sources and years because there are many names for the same species, including multiple scientific names (i.e., synonyms) [4,15]. Perhaps 40% of species names are synonyms and the application of a name can change over time (just as geographic names can) [15,56,69]. This master inventory of species names is critical for molecular- to ecosystem-level studies, but is not yet complete, although progress is being made [56,61,70]. Its completion is feasible, and is a key step in advancing knowledge of life on Earth [4]. Approximately 100 experts have contributed the non-marine components to the Catalogue of Life (CoL [69]), which is at least two-thirds complete, and 200 to the World Register of Marine Species (WoRMS), which is over 95% complete [56,61]. The remaining taxa might be the most difficult to compile, but (given resources) these figures suggest that it should be possible to engage 50–100 new experts to complete the CoL within the next 5 years. The quality assurance and global nature of such a taxonomic inventory will make it the standard and will, in turn, promote further standards in data management. Similarly, molecular databases drove the need for standards to aid data exchange and management that in turn facilitated data analysis and research in genomics and drug discovery (e.g., [71]). Such standards then enable data to be more mobile between databases. For example, WoRMS provides a Webservice that is used by at least 34 organisations [15,56].

Differences between print and data publication

There are differences between conventional print journals and data journals that must be addressed. First, a data set must be published to rigorous domain-specific standards of formatting and structure to enable it to be combined with other data sets, such as by following the 'Darwin Core Archive' data schema [52] so that it can be automatically integrated into GBIF [53] (Box 3). If data users find errors in data, they may inform the data aggregator (e.g., GBIF), who may inform the original data provider. However, this rarely results in the data being corrected at all its locations. Following the more conventional publication model proposed here (Table 1) would reduce the occurrence of such errors, and allow for publication of 'Corrections' and 'Responses'. Second, data sets will often be supplemented by additional data over time. For example, new versions of data sets might have corrected errors and omissions, and time-series data may progressively add new data sets. We agree with others [34,54] that these should be published as new publications because their data will be unique, the authors and metadata can change, and, for time-series data, the temporal scope of

Table 1. A proposed procedure for the publication of biodiversity data sets with a high standard of quality control, including peer review[a]

Process	Quality indicator
1. Online submission of data set for publication with full metadata (title, editors or authors, contact details, abstract, sampling methods, taxa and habitats sampled, keywords, etc.)	★
2. Editor verifies that the data set is within the scope of the journal	
3. Automated tools check data set for omissions and errors, including matching species names against a master list, and mapping geographic data to check against metadata	★★
4. Online tools generate tables of statistics (e.g., how many species per higher taxonomic group, or a species inventory) and maps of data locations	
5. Potential errors and omissions reported to data set author and/or editors	
6. Data set author (or editor as appropriate) responds to report on initial submission technical screening, including resubmitting corrected data and metadata if necessary	
7. Automated data checks verify that data set is complete and standardised. Statistics are recalculated and maps regenerated	
8. Data set author or editor confirms that resubmitted data and metadata are correct	★★★
9. Independent experts (who might be members of an editorial board) assess (i.e., peer review) whether the data set is of sufficient quality for publication	
10. The journal might wish to expose the data set to a wider scientific audience for comments at this time	
11. Author responds to referees' comments	
12. Editor makes a decision on quality standard achieved by the data set, and can ask the data set author or editor to revise the metadata or make other improvements to the data before it will be accepted	
13. Data and metadata are published online having passed several technical checks and peer review. The data set has its own webpage that tracks the metrics of its use. The abstract, citation, authors' contact details, statistics, and maps of the data set will be on this page. The data can be downloaded as tables, comma-separated values, or in other formats as appropriate. Where appropriate, the data set is integrated into the GBIF database, from where it can also be downloaded	★★★★
14. Papers are published that analysed most of the data and any errors found in this process have been corrected	★★★★★

[a]Possible stages where an overall quality indicator can be applied are indicated where five stars is the highest quality. Other quantitative metrics are also recommended (see main text).

each publication will be discrete. However, where changes are limited to minor corrections or amendments, a versioning control system would be preferable.

Steps in data publication

We propose a staged QA/QC process before publication (Table 1). The first stage could use automated tools to test that data are parsable and the geographical coordinates feasible. Such tools have already been developed by GBIF and others [55]. Species names can be automatically checked against authoritative standards (e.g., 'Match Taxa' tool on WoRMS [56]). The next stage could check that metadata are sufficiently informative. A further stage could check the validity of species and place names against a taxonomically authoritative nomenclature and a gazetteer, respectively. A final stage could involve manual peer review by experts, followed by an editorial process and decision similar to research articles. GEON, the Geosciences Network (http://www.geongrid.org) uses a similar strategy for publishing open geological data. Data could be visible online at any stage in the process, but would not have all quality indicators. Thus, the added quality-control options would not impede data publication. Online comment boxes could allow users to comment on published data, and enable data-set authors and editors to provide subsequent information (e.g., announce data additions and new publications that included the data).

Calls for quality metrics in other science fields include a proposed reliability score for mass spectrometry data [57]. Considering the various potential uses of biodiversity data and the principle of fitness for purpose, a 'reliability' index might be difficult to implement. However, measures such as the number of records, species and geographic locations, proportion of species names validated, and data and metadata completeness would be useful to potential users [1,5,16,58]. Examples of such metrics for biodiversity data have been demonstrated by Species Link [59].

Concluding remarks

A new initiative to foster biodiversity data publication is required because the present model cannot cope with the increasing need for availability of high-quality data. Formal publication of data sets, including peer review, is a logical step in scholarly publication and will enable closer integration of publications and databases [60]. Indeed, considering that data underpin information and knowledge, it is at least as important that data sets are peer reviewed as for the papers resulting from their analysis.

Although peer-reviewed data publication might be novel in the new field of biodiversity informatics, it is not radical. Several peer-reviewed journals publish primary environmental and biodiversity data, and primary data have previously been published in monographs, cruise reports, and appendices to papers. We recognise that this standard of quality assurance might not be practical in all situations. Thus, our proposed quality assurance tiers allow data to be published immediately and thereafter be subject to steps of automated, semi-automated, and peer scrutiny (Table 1). Furthermore, instead of these steps being an impediment to data publication, the fact that the final publication will be peer reviewed and published in the style of a conventional scientific journal, will attract scientists whose priority is to 'publish' and for whom 'making data available' is not a priority.

Acknowledgements

We thank Vishwas Chavan, Rod Page, A. Townsend Petersen, Hannu Saarenmaa, Tim Hirsch, Peter Desmet, and the referees for helpful suggestions.

References

1 Costello, M.J. and Vanden Berghe, E. (2006) 'Ocean Biodiversity Informatics' enabling a new era in marine biology research and management. *Mar. Ecol. Prog. Ser.* 316, 203–214
2 Guralnick, R.P. *et al.* (2007) Towards a collaborative, global infrastructure for biodiversity assessment. *Ecol. Lett.* 10, 663–672
3 Wheeler, Q.D. *et al.* (2012) Mapping the biosphere: exploring species to understand the origin, organization and sustainability of biodiversity. *Syst. Biodivers.* 10, 1–20
4 Costello, M.J. *et al.* (2013) Can we name Earth's species before they go extinct? *Science* 339, 413–416
5 Costello, M.J. (2009) Motivation of online data publication. *Bioscience* 59, 418–427
6 Chavan, V.S. *et al.* (2010) *State-of-the-Network 2010: Discovery and Publishing of the Primary Biodiversity Data Through the GBIF Network*. Global Biodiversity Information Facility
7 Global Biodiversity Information Facility (2009) *Participants Report 2009*. Global Biodiversity Information Facility
8 Global Biodiversity Information Facility (2010) *Annual Report 2009*. Global Biodiversity Information Facility
9 Reichman, O.J. *et al.* (2011) Challenges and opportunities of Open Data in ecology. *Science* 331, 703–705
10 Michener, W.K. *et al.* (1997) Nongeospatial metadata for the ecological sciences. *Ecol. Appl.* 7, 330–342
11 Rumble, J., Jr *et al.* (2005) Developing and using standards for data and information in sciences and technology. In *Proceedings of the PV2005 Symposium: Ensuring Long-term Preservation and Adding Value to Scientific and Technical Data*. The Royal Society, Edinburgh, 21–23 November 2005. Available from: (http://www.ukoln.ac.uk/events/pv-2005) (Accessed 3 May, 2013)
12 Baird, R. (2010) Leveraging the fullest potential of scientific collections through digitization. *Biodivers. Inform.* 7, 130–136
13 Parr, C.S. *et al.* (2012) Evolutionary informatics: unifying knowledge about the diversity of life. *Trends Ecol. Evol.* 27, 94–103
14 Page, R.D.M. (2008) Biodiversity informatics: the challenge of linking data and the role of shared identifiers. *Brief. Bioinform.* 9, 345–354
15 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, e51629
16 Chavan, V.S. and Ingwersen, P. (2009) Towards a data publishing framework for primary biodiversity data: challenges and potentials for the biodiversity informatics community. *BMC Bioinformatics* 10 (Suppl. 14), S2
17 Wood, J. *et al.* (2010) *Riding the Wave: How Europe can Gain from the Rising Tide of Scientific Data. Final Report of the High Level Expert Group on Scientific Data*. European Union
18 Piwowar, H.A. *et al.* (2007) Sharing detailed research data is associated with increased citation rate. *PLoS ONE* 2, e308
19 Piwowar, H.A. (2011) Who shares? Who doesn't? Factors associated with openly archiving raw research data. *PLoS ONE* 6, e18657
20 Hampton, S.E. *et al.* (2013) Big data and the future of ecology. *Front. Ecol. Environ.* 11, 156–162
21 Alsheikh-Ali, A.A. *et al.* (2011) Public availability of published research data in high-impact journals. *PLoS ONE* 6, e24357
22 Huang, X. *et al.* (2012) Willing or unwilling to share primary biodiversity data: results and implications of an international survey. *Conserv. Lett.* 5, 399–406
23 Smit, E. (2010) In *Preservation, Access and Re-use of Research Data: the STM View on Publishing Datasets*. Presented at the DataCite Summer Meeting 2010, Hannover, 8 June 2010. Available from: (http://datacite.org/datacite_summer_meeting_2010) (Accessed 3 May, 2013)
24 Thessen, A.E. and Patterson, D.J. (2011) Data issues in the life sciences. *Zookeys* 150, 15–51
25 Chavan, V. and Penev, L. (2011) The data paper: a mechanism to incentivise data publishing in biodiversity science. *BMC Bioinformatics* 12 (Suppl. 15), S2
26 Penev, L. *et al.* (2011) *Pensoft Data Publishing Policies and Guidelines for Biodiversity Data*. Pensoft Publishers
27 Lawrence, B. *et al.* (2011) Citation and peer review of data: moving towards formal data publication. *Int. J. Digit. Curat.* 6, 4–37
28 Santos, C. *et al.* (2005) Supplementary data need to be kept in public repositories. *Nature* 438, 738
29 Vision, T.J. (2010) Open data and the social contract of scientific publishing. *Bioscience* 60, 330–331
30 The Wellcome Trust (2010) *Policy on Data Management and Sharing*. The Wellcome Trust
31 Group on Earth Observations (GEO) (2010) Implementing the Data Sharing Principles. *GEO News* 11
32 Cragin, M.H. *et al.* (2010) Data sharing, small science and institutional repositories. *Philos. Trans. R. Soc. A* 368, 4023–4038
33 Tenopir, C. *et al.* (2011) Data sharing by scientists: practices and perceptions. *PLoS ONE* 6, e21101
34 Altman, M. and King, G. (2007) A proposed standard for the scholarly citation of quantitative data. *D-Lib Mag.* http://dx.doi.org/10.1045/march2007-altman
35 Froese, R. and Pauly, D., eds (2011) *FishBase* (http://www.fishbase.org) (Accessed 3 May, 2013)
36 Guiry, M.D. and Guiry, G.M. (2011) *AlgaeBase* (http://www.algaebase.org) (Accessed 3 May, 2013)
37 Costello, M.J. *et al.* (2007) *About the Ocean Biogeographic Information System*. Available from: (http://hdl.handle.net/2292/5236) (Accessed 3 May, 2013)
38 Ware, M. and Mabe, M. (2009) *The STM Report: An Overview of Scientific and Scholarly Journal Publishing*. International Association of Scientific, Technical and Medical Publishers
39 Abbott, A. *et al.* (2010) Do metrics matter? *Nature* 465, 860–862
40 Parsons, M.A. *et al.* (2010) Data citation and peer review. *EoS* 91, 297–299
41 Chavan, V. *et al.* (2013) Cultural change in data publishing is essential. *Bioscience* (in press)
42 Sieber, J.E. and Trumbo, B.E. (1995) (Not) giving credit where credits due: citation of data sets. *Sci. Eng. Ethics* 1, 11–20
43 Birney, E. *et al.* (2009) Prepublication data sharing. *Nature* 461, 168–170
44 Constable, H. *et al.* (2010) VertNet: a new model for biodiversity data sharing. *PLoS Biol.* 8, e1000309
45 Mons, B. *et al.* (2011) The value of data. *Nat. Genet.* 43, 281–283
46 Whitlock, M.C. (2011) Data archiving in ecology and evolution: best practices. *Trends Ecol. Evol.* 26, 61–65
47 Costello, M.J. *et al.* (2012) *Quality Assurance and Intellectual Property Rights in Advancing Biodiversity Data Publications*. (ver. 1.0), Global Biodiversity Information Facility
48 Chavan, V. (2012) *Recommended Practices for Citation of the Data Published through the GBIF Network*. (ver. 1.0), Global Biodiversity Information Facility
49 Page, R.D.M. (2006) Taxonomic names, metadata, and the semantic web. *Biodivers. Inform.* 3, 1–15
50 Gahegan, M. *et al.* (2009) Connecting GEON: making sense of the myriad resources, researchers and concepts that comprise a geoscience cyberinfrastructure. *Comput. Geosci.* 35, 836–854
51 Penev, L. *et al.* (2009) Publication and dissemination of datasets in taxonomy: Zookeys working example. *Zookeys* 11, 1–8
52 Wieczorek, J. *et al.* (2009) *Darwin Core*. Available from: (http://www.tdwg.org/standards/450) (Accessed 3 May, 2013)
53 Wieczorek, J. *et al.* (2012) Darwin Core: an evolving community-developed biodiversity data standard. *PLoS ONE* 7, e29715
54 Klump, J. *et al.* (2006) Data publication in the open access initiative. *Data Sci. J.* 5, 79–83
55 Harnad, S. (2008) Validating research performance metrics against peer rankings. *Ethics Sci. Environ. Politics* 8, 103–107
56 Appeltans, W. *et al.* (2012) *World Register of Marine Species*. Available from: (http://www.marinespecies.org) (Accessed 3 May, 2013)
57 Gough, N.R. and Yaffe, M.B. (2011) Focus issue: conquering the data mountain. *Sci. Signal.* 4, 1–3
58 Heidorn, P.B. (2008) Shedding light on the dark data in the long tail of science. *Libr. Trends* 57, 280–299
59 Centro de Referência em Informação Ambiental (CRIA) (2011) *Species Link*. Available from: (http://splink.cria.org.br/dc/index?&system=&setlang=en) (Accessed 3 May, 2013)

60 Bourne, P. (2005) Will a biological database be different from a biological journal? *PLoS Comp. Biol.* 1, e34
61 Appeltans, W. *et al.* (2012) The magnitude of global marine species diversity. *Curr. Biol.* 22, 1–14
62 Costello, M.J. *et al.* (2012) Predicting total global species richness using rates of species description and estimates of taxonomic effort. *Syst. Biol.* 61, 871–883
63 Saeedi, H. and Costello, M.J. (2012) Aspects of global distribution of six marine bivalve mollusc families. In *Clam Fisheries and Aquaculture* (da Costa, F., ed.), pp. 27–44, Nova Science Publishers
64 Peterson, A.T. *et al.* (2010) *Forward-looking Report*. Global Biodiversity Information Facility
65 Global Biodiversity Information Facility Secretariat (2011) *Publications that cite GBIF*. Available from: (http://www.editgrid.com/user/gbif_secretariat/Professional_Publications_that_cite_GBIF) (Accessed 3 May, 2013)
66 Gargouri, Y. *et al.* (2010) Self-selected or mandated, open access increases citation impact for higher quality research. *PLoS ONE* 5, e13636
67 SQW Ltd. (2004) *Costs and Business Models in Scientific Research Publishing: A Report Commissioned by The Wellcome Trust*. The Wellcome Trust
68 Houghton, J. *et al.* (2009) *Economic Implications of Alternative Scholarly Publishing Models: Exploring the Costs and Benefits*. JICS
69 Costello, M.J. *et al.* (2013) More taxonomists describing significantly fewer species per unit effort may indicate that most species have been discovered. *Sys. Biol.* http://dx.doi.org/10.1093/sysbio/syt024
70 Bisby, F.A. *et al.*, eds (2012) *Species 2000 & ITIS Catalogue of Life: 2011 Annual Checklist,* Species 2000
71 The Gene Ontology Consortium (2000) Gene ontology: tool for the unification of biology. *Nat. Gen.* 25, 25–29