Quality assurance and Intellectual Property Rights in advancing biodiversity data publication

Version 1.0

October 2012

Suggested citation:
Mark J Costello, William K Michener, Mark Gahegan, Zhi-Qiang Zhang, Phil Bourne, Vishwas Chavan (2012). Quality assurance and intellectual property rights in advancing biodiversity data publications ver. 1.0, Copenhagen: Global Biodiversity Information Facility, Pp. 33, ISBN: 87-92020-49-6. Accessible at http://links.gbif.org/qa_ipr_advancing_biodiversity_data_publishing_en_v1.

ISBN: 87-92020-49-6
Persistent URI: http://links.gbif.org/ qa_ ipr_ advancing_biodiversity_data _publishing_en_v1

Language: English

Project Partners: The Global Biodiversity Information Facility (GBIF

Document Control:

Version	Description	Date of release	Author(s)
0.8	Content development	January 2012	Mark J Costello, William K Michener, Mark Gahegan, Zhi-Qiang Zhang, Phil Bourne, Vishwas Chavan
0.9	Review, edits	April 2012	Mark J Costello, William K Michener, Mark Gahegan, Zhi-Qiang Zhang, Phil Bourne, Vishwas Chavan
1.0	Final version	October 26 2012	Mark J Costello, William K Michener, Mark Gahegan, Zhi-Qiang Zhang, Phil Bourne, Vishwas Chavan

About GBIF

GBIF: The Global Biodiversity Information Facility

GBIF was established by countries as a global mega-science initiative to address one of the great challenges of the 21st century – harnessing knowledge of the Earth's biological diversity. GBIF envisions 'a world in which biodiversity information is freely and universally available for science, society, and a sustainable future'. GBIF's mission is to be the foremost global resource for biodiversity information, and engender smart solutions for environmental and human well-being (GBIF 2011a). To achieve this mission, GBIF encourages a wide variety of biodiversity data holders, generators and users across the globe to discover and publish (make discoverable) data to global standards through the GBIF network. Website: http://www.gbif.org.

Table of Contents

Contents

List of Figures

List of Boxes

List of Tables

Executive Summary

An unprecedented amount of biodiversity data is becoming available on the internet. However, significant amounts of data, particularly historic data, are not available online. The Global Biodiversity Information Facility publishes millions more primary biodiversity data records every year, but finds that this is a decreasing proportion of the potentially available data it could publish. Because data sharing agreements and policies alone are insufficient, new approaches are required to accelerate data publication. Only in the past few years have scientists begun calling for data 'citation' and referring to data 'publication' rather than data 'sharing' and 'availability'. Issues of intellectual property rights (IPR) only complicate data access in the latter contexts. In contrast, the 'publication' process has well-established conventions that simplify and clarify IPR issues.

Concerns over data quality impede the use of large biodiversity databases by researchers and subsequent benefits to society. Peer-review is the standard mechanism used to distinguish the quality of scientific publications. Here, we argue that the next step in data publication is to include the option of peer-review. Data publication can be similar to the conventional publication of articles in journals that includes online submission, quality checks, peer-review, and editorial decisions. This quality-assurance process will at least assess, and potentially could improve the accuracy of the data, which in turn reduces the need for users to 'clean' the data, and thus increases data use while the authors and/or editors get due credit for a peer-reviewed (data) publication. Adoption of international and community-wide standards related to data citation, accessibility, metadata, and quality control would enable easier integration of data across datasets. Metadata, for example, would include relevant information about the datasets that would enable a user to better understand the data and determine its suitability for use for particular purposes.

It is recognized that a significant amount of data is already published without peer-review, both through GBIF and other databases, and through various internet and print media. This will continue. However, providing a scale of quality assurance, of which the highest standard is peer-review, will both improve quality assurance and attract the attention of scientists and organizations that place little value on non peer-reviewed publications. Most steps in the process proposed here are already undertaken by GBIF and/or some of their participants. The peer-review process is well-established in the science community, including peer-review of biodiversity data by several journals. Thus the process proposed here is practical and does not pose new technical difficulties. It may be implemented by GBIF in collaboration with its participants and science journals.

Data publications should strive to be of similar merit as other peer-reviewed publications, and thus be recognized by employers, funding agencies and scientists as a meritorious activity. This will require metrics of data use, such as views, downloads and citations. Here, we propose a staged publication process involving editorial and technical quality controls, of which the final (and optional) stage includes peer-review.

Section 1: Introduction

In today's digital world, all biodiversity information and data should be available online, unless there are sound reasons why they be kept confidential (e.g. rare bird nesting site). Information that is not online will be overlooked by most readers. For biodiversity data the requisite storage capacity and infrastructure are available, and there are continuing improvements in indexing and automated tools for data management (e.g. Costello and Vanden Berghe 2006, Guralnick et al. 2007). However, quality assurance is inconsistent and a culture of data publication is lacking. Consequently, relatively few scientists use biodiversity databases for their research, and few scientists contribute data back to the community. While millions of dollars of important, publicly-funded data are 'lost', global issues remain such as climate change, over-fishing, infectious diseases, and invasive species, threatening human food sources and ecosystem health. Addressing these challenges requires that existing data be published, properly maintained and openly accessible.

Biodiversity data can include inventories of species names and their synonyms, data on species distributions in one or more places and times, images and sounds of the species or their anatomy, ecological interactions, behaviour, descriptions of the dataset, and analyses and interpretations of the data (Costello 2009a). In this paper we are most concerned with the primary biodiversity data rather than the secondary (e.g. modelled or simulated) data derived from it, and interpretations and descriptions around data. Data may thus be numerical, categorical (e.g. species or place names), images or sounds. Examples of datasets include: bird counts; insects captured by light and pitfall traps, or canopy fogging; fishery trawl data; benthic macro-invertebrate surveys; counts and images of whale observations; water quality monitoring that includes biological indicator species; specimen collections (e.g. in museums); results of ecological research studies; habitat and biotope maps; and compilations of data from the literature.

Making data available increases visibility of scientists' work, and can increase the citations of their papers (e.g. by 70% for cancer clinical trials, Piwowar et al. 2007). This may be an incentive to some scientists, but still less than half of authors make their data publicly available online (Piwowar et al. 2007, Piwowar 2011). Even in those journals which have a policy that data should be made available, one study found most (59%) papers did not follow it (Alsheikh-Ali et al. 2011). Another survey found that while 80% of scientists wanted access to data created by others, only 13% did not want to share their data, but only 20% have actually shared data (Smit 2010). Clearly, data sharing agreements and

policies are insufficient, and new approaches are required to increase the availability of high quality biodiversity data (Costello 2009a).

Available data are increasing

Never before have there been so many scientists publishing so many papers and books (Ware and Mabe 2009). There are over 26,000 scholarly journals at present, and about 50 million research articles published by the year 2010 (Jinha 2010). Advances in technologies, from medical studies to mobile phones, satellites to environmental sensors and videos, enable ever-increasing amounts of digitized data to be automatically captured (Porter et al. 2011, Michener and Jones 2011). The Global Biodiversity Information Facility (GBIF) publishes millions more records of species every year (Chavan et al. 2010). However, centuries of irreplaceable historic data on biodiversity and the environment need to be moved into the digital environment to provide the historical context for present observations, and enable predictive modelling of the consequences of human activities for the environment and biodiversity, including human food and ecosystem services (Michener et al. 1997). These past observations cannot be reproduced and thus must be a priority for digitization (Rumble et al. 2005, Baird 2010, Parr et al. 2012). This historic record is especially important for taxonomy, because the first description of a species has legal priority for that species name (Page 2008). In contrast, biomedical data are largely a recent phenomenon.

Data centres and repositories do exist but this does not necessarily mean that there is sufficient motivation for scientists or organizations to submit data to them. Despite the large volume of data published through GBIF, significant amounts of historic data are not yet included and more data are continually being collected (Yesson et al. 2007, GBIF 2009, 2010, Reichman et al. 2011). There is a need to motivate and reward the contribution of data to international integrated databases by bringing data publication into the mainstream of respected scientific publications (Costello 2009a, Chavan and Ingwersen 2009, Wood et al. 2010, Moritz et al. 2011, Reichman et al. 2011,).

Need for standards

A benefit of integrating data into one system is that it drives standards for data management. Standardized, quality assured, permanently archived databases are essential to manage the collection, storage and accessibility of this growing data stream. This is widely recognized (e.g. Wood et al. 2010), 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). Over 20% of species names are synonyms and the application of a name can change over time (just as geographic names can) (e.g. Gaston and Mound 1993, Chavan et al. 2005, Stork et al. 2008). This master-inventory of species names is critical for molecular to ecosystem level studies but is not yet complete, although progress is being made by the taxonomic community (e.g. Bisby et al. 2010, Appeltans et al. 2012). Its completion is feasible. About 100 experts have contributed the non-marine components to the Catalogue of Life (CoL, Bisby et al. 2011) which is at least two-thirds complete, and 200 to the World Register of Marine Species (WoRMS) (Costello and Appeltans 2008, Appeltans et al. 2012) which is over 90% complete. The remaining taxa may 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 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. The Gene Ontology Consortium 2000).

Section 2: Data Publication

Decades ago, it was not uncommon for journals and monographs to publish species inventories, data from ecological surveys, and data appendices. However, the cost of print and postage led to journals being very reluctant to publish tables and appendices of primary data. Today, the availability of online appendices and electronic publication means this is no longer an issue, and at least some biodiversity journals, e.g. *Zootaxa* and *Phytotaxa,* publish species inventories both in print and online, and the European Register of Marine Species (ERMS, Costello 2001) database was published as a special issue of a journal and a book (Costello et al. 2000). Thus the concept of publishing primary biodiversity data is not new.

There appears unanimity amongst inter-governmental organizations, governments, science journals and science funding agencies that at least data created using public funds or for the public good (e.g. environmental monitoring data) should be publicly available

(Costello 2009a, Thessen and Patterson 2011, Chavan and Penev 2011). To be made 'public' implies 'publication' (Box 1). Scientific publishers increasingly expect authors of papers to make their data publicly available, ideally in international databases, in permanent institutional repositories, or as online supplementary material (reviewed in Costello 2009a). However, the peer-review and editorial processes generally exclude assessment of the associated data. Important exceptions include the Ecological Society of America's (ESA) *Data Papers* and *Ecological Monographs*, the *Earth System Science Data Journal* (ESSD), and *Marine Biodiversity Records*. A new open-access journal, *Aquatic Invasions,* publishes peer-reviewed papers with distribution data on species and has recently established a sister journal, *BioInvasions Records* for data papers (Panov et al. 2011). This year, another new journal *Datasets in Ecology* (www.datasets.com) was launched, and the publisher Pensoft has announced the introduction of 'data papers' in six of its journals. Thus unless the authors publish in a specialist 'data journal', there is no oversight to ensure the dataset is to an international standard, has adequate metadata, and is largely free from errors.

Online appendices to printed scientific papers are not an ideal method of publication because they are not necessarily peer-reviewed (Lawrence et al. 2011), and may not be subjected to independent editorial attention. Because such appendices are not required to conform to standards for data or metadata, their re-use can be problematic. Furthermore, a significant portion of such 'supplemental materials' become inaccessible over time (Santos et al. 2005, Vision 2010). Institutional repositories may be preferable where they provide permanent archiving, but most lack peer-review, editorial-review and the ability to be familiar with emerging standards for all disciplines. Dryad provides the option for authors of papers published in biodiversity related journals to deposit their datasets in a central open-access repository (http://datadryad.org). By early 2012, it had published over 3,000 data files from articles published in 100 journals. Specialized data centres are most familiar with data standards, and in-house staff provide some quality assurance of data and metadata quality (e.g. PANGAEA and the Distributed Active Archive Centers (DAAC's) of the National Aeronautics and Space Administration (NASA)), However, only GBIF and similar organizations, including GBIF participants, provide this specialist attention to integrate biodiversity data. Thus, specialized data centres and databases are preferable for data publication. These include GenBank, Protein Data Bank and similar systems that manage molecular and chemical data. These databases provide unique

services to academic, governmental and commercial researchers. However, biodiversity and environmental data have lagged behind, and the concept of data publication is not a widespread practice within biodiversity science.

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' them (e.g. Wellcome Trust 2010, Group on Earth Observations 2010). This is a key distinction, because making something available suggests a negotiation between the parties involved as to the terms and conditions of data availability. This may require direct payment, joint authorship of scientific papers for which the data are used, or partnership in research contracts (e.g. Costello 2009a, Cragin et al. 2010, Tenopir et al. 2011, Thessen and Patterson 2011). Fortunately, this is not the case for scientific papers, and should also not be for datasets (Altman and King 2007). These calls for making data 'available' may be counter-productive because they pressure scientists to do something outside their comfort zone. They may not have clarified data ownership and dissemination policy with their collaborators, employer, and/or funding sources. By giving their data away they may be criticized by these organizations and/or individuals for having handed competitors an advantage, and compromised their ability to re-use the data and/or leverage more funding based on it. Furthermore, a significant amount of work may be required to get the data into a well-described format that others can use. Whether or not 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. In contrast, 'publication' is the normal expectation of funding agencies, collaborators, and employers of researchers.

Biodiversity data publication

There are a suite of open-access scholarly biodiversity databases on the World Wide Web (e.g. listed in Thessen and Patterson 2011). Most provide information on species, with well-established examples such as FishBase (Froese and Pauly 2011), AlgaeBase (Guiry and Guiry 2011), and the Global Invasive Species Database (Anon. 2011a), and emerging systems using images, such as to identify individual whales from their photographs (www.cetabase.info). A few provide standardized distribution data, such as VertNet (Constable et al. 2010) and the Ocean Biogeographic Information System (OBIS, Costello et al. 2007), and feed (and republish) data into GBIF. GBIF was established to make biodiversity data publicly available and thus satisfy a key aim of the Convention on Biological Diversity. As an inter-governmental organization financially supported by -

participant countries it is by far the largest resource for biodiversity data in the world and the only one with an inter-governmental governance and funding structure. GBIF and OBIS are organized as global networks of national, regional and thematic nodes that compile and deliver data to a database which is published through a single portal. In contrast, some molecular databases began as competing initiatives that now exchange data but remain independently managed and thus retain an element of competition (e.g. GenBank). Although this paper focuses on the role of GBIF, its principles can be applied to other biodiversity data initiatives.

Over its first decade, GBIF published over 300 million records of species, from 14,000 datasets supplied by 400 organizations from over 40 countries, with over 4.5 million names (Figure 1). The names include scientific, vernacular and other names, and amounts to almost 1 million species of which 590,000 have distribution data (Tim Robertson, pers. com. 4th July 2012). The marine component of GBIF, OBIS, contains over 120,000 species which is over half of all described marine species (Costello et al. 2011). Similarly, searches of the occurrence of bivalve molluscs found that about half the known species have data in OBIS and GBIF (Saeedi and Costello, 2012). Despite a widespread impression that museum specimens comprise most of the data, about 80% represent species observations and samples (GBIF 2010). The data from each source are organized in standardized tables so they can all be integrated into a large searchable database (Wieczorek et al. 2012). Most data are georeferenced, so that over 85% of animals and 76% of plant species can be mapped (Chavan et al. 2010). The sum of local and regional data can thus be used to examine global scale phenomena. Over two-thirds of the datasets in GBIF have been provided by government organizations whose staff are directed to do so. Far fewer datasets are delivered from the academic community although it publishes *ca.* 75% of all scientific papers, despite comprising only 15–50% of all scientists (Ware and Mabe 2009).

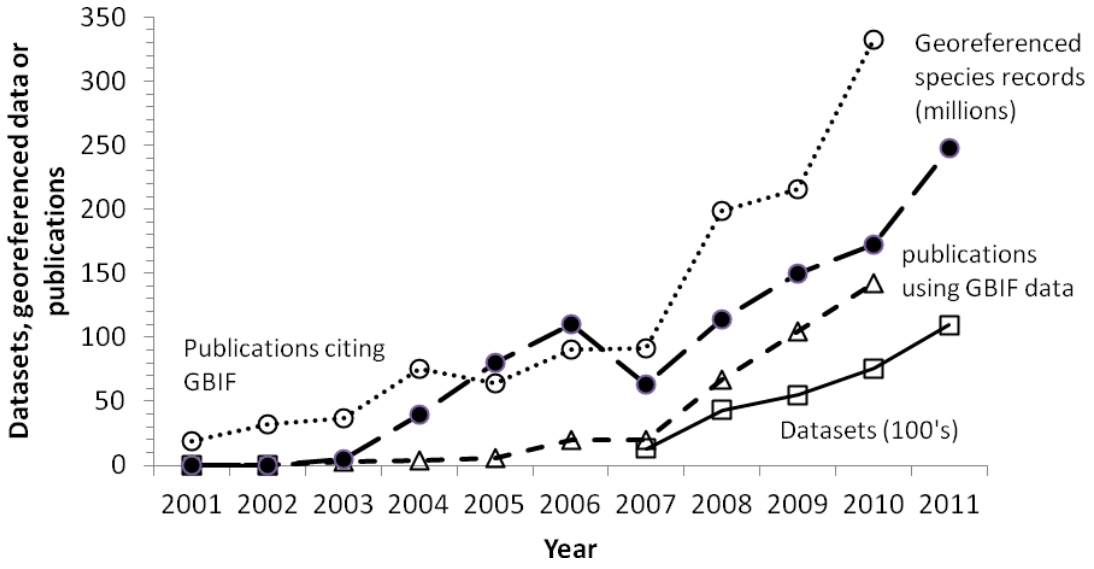

Figure 1. The number of (a) millions of species distribution records published by GBIF (solid dots), (b) hundreds of datasets (squares), (c) publications that use data from GBIF (triangles), and (c) publications that reference GBIF (hollow circles).

Although the quantity of published records in GBIF has been increasing, the amount of biodiversity data not available to it is also increasing (Figure 2). GBIF has been capturing a decreasing proportion of the biodiversity data that are available (Chavan et al. 2010). This publication deficit needs to be addressed for GBIF to be more complete. Additionally, the spatial and temporal coverage of GBIF data is very uneven. A critique of GBIF progress recognized that criticisms by scientists over data accuracy may be impeding data reuse and consequent benefits to society and thus recommended a greater focus on data quality (Peterson et al. 2010). Nevertheless, the number of publications that used data from GBIF is increasing (Figure 1). Thus issues of both GBIF data completeness (quantity) and accuracy (quality) must be urgently addressed. GBIF needs to address not just the amount of data, but geographic, temporal and taxonomic coverage.

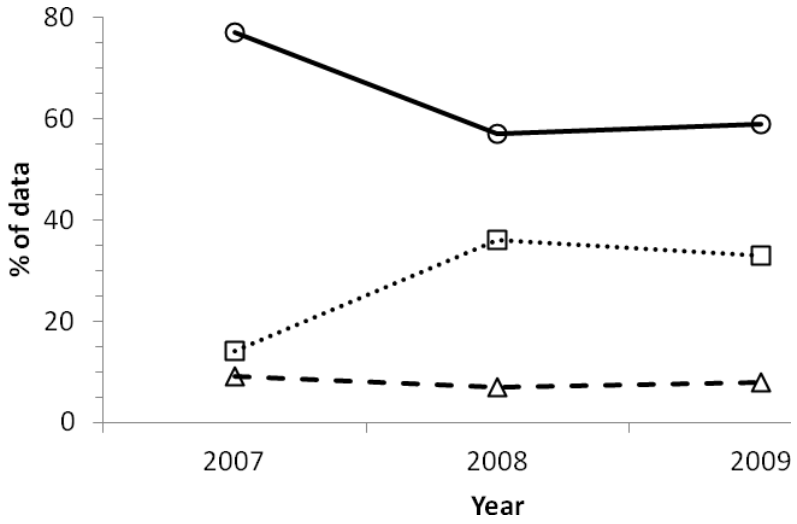

Figure 2. The percentage of data provided to GBIF (circles) compared to the amount of species distribution records digitised but not provided to GBIF (squares), and not digitised biodiversity data (triangles) as reported by the GBIF community in 2007 (n = 33 respondents), 2008 (n = 29) and 2009 (n = 27).

The process of biodiversity data publication through GBIF is that member countries and organizations send data, or they endorse others to send datasets to GBIF. The data are submitted or linked to GBIF in standardized files which are then made available online. Datasets are described by 'metadata' that include the source, version, and how they should be cited. However, the data are not subject to prior approval or review by anyone except the scientists who delivered and/or created them. This process may have been acceptable in the early days of GBIF but such a self-evaluation process is not one that scholarly science journals follow.

Initially GBIF and OBIS proposed an entirely distributed data system. Users would enter search terms at the portal and it would search a network of nodes for responses. Thus data providers could control what data was being released and stop the release of their data if they wished. However, as the system grew searches would slow, and unless the entire content was analysed and indexed, the portal would not know what data was potentially available. A 'no response' could mean no data was available or one node was offline or some other temporary error. Now, like most search engines, GBIF caches all data, effectively creating a centralized database enabling faster system responses and more accurate searches by for example, species, higher taxon, and/or geographic

location. Thus, the data providers are no longer in direct control of their data and GBIF has become the 'publisher'. Indeed, it is doubtful that once made available on the internet the data could ever be withdrawn even if GBIF no longer published it; it may well have been downloaded by users. Our description of GBIF as the publisher of the data thus helps clarify the reality of the situation and may prompt some reflection as to the language used around this process within GBIF.

Data publication through GBIF is opportunistic rather than strategic or aimed to fill gaps demanded by users (Berents et al. 2010). A more incentivized publication model may encourage scientists to offer datasets to GBIF for publication, just as they now offer papers to journals to publish. This may 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 may be more attractive to scientists and data users.

Intellectual property

Although individual data items or facts are not copyrightable (Box 1), compilations of data into checklists or other contexts usually are because they involved some organization, design and intellectual activity in their creation (Van den Eynden et al. 2011). Field sampling programmes will have involved consideration of sampling design and methods, and subsequent laboratory analyses and data processing.

In addition to the minimal incentives to encourage scientists to publish their data, the use of terms such as data exchange or sharing, and consequent opaqueness about copyright and licensing arrangements do not help. Issues of intellectual property, data, database and dataset ownership, and licensing agreements for data use arise when data are 'shared' or 'made available' (Box 1). These terms imply some negotiation or agreement. Consequently, a quagmire of different policies, national legislation regarding copyright and data protection, expectations of users, and different licences for commercial and non-commercial users, surface and effectively stifle data access (Costello 2009a, Tenopir et al. 2011). However, similar issues affect conventional scientific publication, but the only expectation is that the publications are appropriately cited. Of course access to conventional publications usually requires the reader or library to purchase the book, subscribe to the periodical, or for the author or a sponsor to pay for it to be 'open-access'. There are some exceptions, such as when an institution or government funds the publication process. Such an example is the open-access online *European Journal of*

Taxonomy where neither the author nor readers pays (Bénichou et al. 2012). The extent to which this institutional financial support can increase in the long term may dictate the number of articles and pages the journal can ultimately produce. In contrast, author-pay and reader-pay models can achieve economies of scale as the journals grow in output.

Data portals that publish data from centralized or distributed sources should have a common policy, such as a Creative Commons license, for the use of their integrated datasets (Hagedorn et al. 2011, Desmet 2012). This is independent of what person or organization holds copyright on any of the data. Science journals and book publishers may or may not hold copyright on the material they publish, but they provide it for use in a consistent way. By allowing each dataset to have its own license terms in the dataset metadata, GBIF has different terms of use for different datasets. However, this information is not easily visible and will thus be overlooked by most users who may then inadvertently contravene some license conditions. This does not mean scientists need to sign over data ownership (GBIF does not assert any copyright over the data), but they implicitly give GBIF the right to index, clean, and publish the data integrated with other data so that they conform to a common standard and recognize this added value. If there are reasons not to make some data publicly available (e.g. commercial or conservation sensitivity), then it should not be published. This model can also apply to open-access journals. For example, the BioMed Central and Public Library of Science journals let authors retain copyright on the content of their papers (Bourne 2005).

A survey of GBIF participants explored IPR issues that impeded publication (GBIF 2009). It found 60% of respondents felt the current IPR framework was adequate and none found it inadequate. A survey in 2011 of how GBIF participants addressed IPR on their own websites found 51% (n = 35) had statements requiring compliance with 'Terms and Conditions' of use, but usually these were not visible. Surprisingly, 26% had no copyright statement, 40% had no statement indicating they expected to be credited, and 70% did not provide guidelines on how the resource should be cited. This illustrates how the scientific community in general is unclear about how to cite and credit online resources. However, FishBase (Froese and Pauly 2011), AlgaeBase (Guiry and Guiry 2011), MycoBank (Robert et al. 2005), and Index Herbariorum (Thiers 2012), provide recommended citations on their web site, and WoRMS does so at both a database and web page level (Appeltans et al. 2012). Thus it appears that IPR issues are not necessarily a constraint in publishing data online. However, GBIF participants did feel that incentives for data publication through GBIF were insufficient (GBIF 2009). In particular, clearer citation of datasets, use of

citations to give credit to the author or editors, a citation index to track usage (e.g. downloads, searches), and other methods to improve the career impact from publishing data were required. Yet most datasets in GBIF have not been accompanied by sufficient metadata to enable citation in a conventional manner. Here, we propose that presenting data publication within the well-established framework of scientific publications can help resolve these issues. This should include making familiar conventions for citation (as in the Reference list of this paper) and use (e.g. Creative Commons options) easily visible.

Section 3: Quality Assurance and Control

It is common for datasets to be accompanied by statements about the publisher and/or creators not being responsible for the use others may make of the data, or for any errors contained in a dataset. In a 2011 survey of GBIF participants, none used positive statements about the quality or completeness of their data. Of 35 respondents, 57% stated they could not guarantee the quality of the data and 43% had no statement about data quality. A similar proportion (54%) said they were not responsible for the accuracy and reliability of the data; which begs the question as to who could be. Such disclaimers are not prominent in conventional scientific journals. Surely the authors of papers and their publishers have responsibilities to ensure the publication is of good quality. This often includes a prior review of the submitted paper by editors and independent experts (i.e. peer-review). Furthermore, if after publication some errors, plagiarism, or other defects are found then they are corrected, or in extreme cases, the publication may 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. Thus, instead of having a disclaimer, data publishers should be proactive and use transparent QA/QC procedures like scientific journals do.

There are several components to quality assurance. There should be sufficient information about the dataset (i.e. metadata) to indicate its creator and content, and thus fitness for a particular purpose. We propose this include the conventional components of a 'citation' used in the print media, i.e. author (creator or editor), title indicating its content, and source; as recently recommended by GBIF (2012). Furthermore, by classifying and 'indexing' the actual data, it becomes possible for users to also search and select

particular data from several datasets directly, and later study the metadata to judge its suitability.

One strategy is to adopt a staged QA/QC process prior to publication, since different users and organizations may have varying constraints on the time that may be devoted to quality control processes (Table 1). The simplest stage might be that the data are parsable and the geographical coordinates feasible — an automatable process. The next stage may involve the validity of the place and species names used. A further stage could involve manual peer review by experts, followed by an editorial process and decision similar to research articles. Following publication, various further reviews are possible based on feedback from users. With such a staged approach, researchers or organizations that cannot commit time to resolving issues with the data are not precluded, though of course the system and the community need to understand to what stage the QA/QC process has been carried out. GEON, the Geosciences Network (www.geongrid.org) uses a similar strategy for the publication of open geological data.

Fitness for purpose

Understandably, in the early stages of collating data the emphasis is on the quantity of data. For example, in its early years the Protein Data Bank (PDB) had to spend one-third of its budget on data cleaning, a process only begun for some biodiversity databases. Considering that all the purposes to which data may be put are unlikely to be predicted, and most data may be useful for some purposes, then there can never be too much biodiversity data. However, a large amount of data may be insufficient for one purpose, while very little data may be adequate for another. Establishing that data are fit for purpose is often a very difficult task, and may entail a study of both metadata, and the processes (workflow) used to create the data, as well as their content. Enhancements to metadata should be driven by the need to help users understand what they could and could not use the data for. An example of such metadata, developed for the characterization of datasets published by OBIS, is appended (Appendix).

GBIF has been operational for over a decade, and its role is to provide access to biodiversity data. There may be an intrinsic value in compiling data into databases because it provides future opportunities for data analyses by others in the scientific community, including future generations. However, in a world of competing demands for funds and personnel time, it is hard to prioritize this option without demonstrated uses of the data. To convince people of this value, these demonstrations must provide new

insights only possible through the creation of the database. Fortunately an increasing number of such papers are now being published (Figure 1).

Table 1. A proposed procedure for the publication of biodiversity datasets with a high standard of quality control, including peer-review. Possible stages where an overall quality indicator can be applied are indicated. More quantitative metrics are also recommended (see text).

Process	Quality indicator
1. Online submission of dataset for publication with full metadata (title, editors or authors, contact details, abstract, sampling methods, taxa and habitats sampled, keywords, etc.).	★
2. Editor verifies the dataset is within the scope of the journal.	
3. Automated tools check dataset 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, inventory of species) and maps of data locations.	
5. Potential errors and omissions reported to dataset author and/or editors.	
6. Dataset 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 dataset complete and standardized. Statistics are recalculated and maps re-generated.	
8. Dataset author or editor confirms re-submitted data and metadata are correct.	★★★
9. Independent experts (who may be members of an editorial board) assess (i.e. peer-review) whether the dataset is of sufficient quality for publication.	
10. The journal may wish to expose the dataset 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 dataset, and may ask the dataset author or editor to revise the metadata or make other improvements to the data before it will be accepted.	
13. Data and metadata published online having passed several technical checks and peer-review. The dataset has its own webpage that tracks the metrics of its use. The abstract, citation, authors' contact details, statistics and maps of the dataset will be on this page. The data can be downloaded as tables, comma separated values, or in other formats as appropriate. Where appropriate, the dataset is integrated into the GBIF database and can also be downloaded from there.	★★★★
14. Papers are published that analyzed most of the data and any errors found in this process have been corrected.	★★★★★

Dataset citation

Smit (2010) found 92% of scientists wanted credit for the use of their data, as would be provided by citation of a data publication or mention in the Acknowledgements of a paper. In the process proposed here, the editors will determine the citation style for their journal, but one can expect it to include the common elements of authors, title, publisher, and date of publication (Altman and King 2007, Costello 2009a, Parsons et al. 2010). Costello (2009a) listed 16 benefits of data publication, but 9 of these can only be realized if the data are cited in this way. Following an established publication process implies standard citation of datasets, tracking of citations, and other metrics of their use (e.g. page views, downloads) (Costello and Vanden Berghe 2006, Chavan and Ingwerson 2009, Costello 2009a, Ingwersen and Chavan 2011). A further expectation of a cited publication is that it will be permanently accessible as published, even if the dataset is modified or expanded in the future.

At present, journals and authors have different policies on how to cite online resources. Some include a url in the paper text, rather than citing them 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 may change over time (Altman and King 2007). The publication of datasets in a more conventional manner, that is like '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 datasets it published in 2006 (Appendix) and a variety of options have been considered by GBIF (2012).

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 Life Science Identifiers (LSID), Digital Object Identifiers (DOI), Uniform Resource Names (URN), or similar unique identifiers that aid databases in tracking citations of publications (e.g. Page 2006). In addition to tracking citation of datasets, there is also the opportunity to track the use and provenance of individual items of data (Page 2008). This provides new opportunities to develop metrics of data use that could be used to recognize the impact of data publication. The ability to track data views and downloads already exists and has been implemented by some journals. A range of data use metrics are necessary because datasets may not always be cited and tracked by scientific abstracting services (Costello and Vanden Berghe 2006).

When a data paper links to the source of a data set or database, typically a url (universal resource locator) is included. However, a url can change when data sets are moved or domain names are changed. What would be needed for data publication is a central url registry for data sets (similar to what CrossRef is doing for papers) so that when the url for data sets change, the records in the registry are updated. This will ensure that links to data sets are persistent. CrossRef registers DOIs for papers published by member publishers. When the url for journals or papers are changed, CrossRef records are updated so that the DOIs cited in the references always point to the correct urls for registered papers. Various digital object identifiers are already being assigned to data sets. For example, DataCite (Anon. 2011b) and Dryad are already providing DOIs to datasets, whereas other organizations are assigning LSIDs to data sets. GBIF and other major data publishers should take a lead in developing such a central registry for resolving various digital identifiers for permanent linking to the correct url for data sets.

Data archiving

The persistence of data sets is even more important than ensuring the correct link to online data via a central registry. Major biodiversity dataset repositories should work together to enable authors of data sets to deposit and archive their data more easily. They should provide free or low cost assignments of persistent identifiers which can be centrally registered. Any new data journals should work closely with these repositories and recommend that authors use creditable repositories (e.g. GBIF for georeferenced distribution data, GenBank for gene sequence data, MorphoBank for phylogentic data matrix etc). Storing datasets with individual journal websites should be discouraged, because individual journals and publishers may discontinue and, more importantly, aggregated storage of large amounts of similar data sets can greatly facilitate their discovery and use for new purposes.

Indexing

Problems in taxonomic nomenclature and geographic accuracy are well known (e.g. Yesson et al. 2007, Hill et al. 2010). They can be identified and partly addressed by semi-automated quality control checks as part of a data indexing process. In biodiversity databases, this requires classification of species into higher taxonomic groups (or taxa), matching synonyms to their species, mapping the geographic locations of records, and recording when they were collected, and who collected them for what purpose. Comparison of the results of such indexing of the data with the description of the dataset

by its authors or editors, also called 'descriptive or discovery metadata', is a first step in quality assurance. For example, it should find that all the species expected are accounted for, and that the data map to the right place. All this information and more are collected in a standardized form by GBIF, enabling ever-improving abilities to search the database on these criteria, including generating species inventories, mapping species and predicting their ranges using environmental data. The outcome of the above technical 'cleaning' of the data could be used to assess the quality of datasets, as demonstrated by SpeciesLink (CRIA 2011), and provide statistics and other outputs that would assist experts in assessing dataset quality. This process will be more demanding of data creators, but justified by the more prestigious nature of the subsequent quality assured publication. The completeness of data and metadata in a dataset could provide an index of fitness for particular purposes (Michener et al. 1997). For example, not all species data published through GBIF contains the species location, and not all datasets contain information about the species habitat or sampling method.

Peer review

Despite the time-consuming nature of the peer-review process, it is widely regarded as an essential part of science (Ware and Mabe 2009). Publications without this quality control are regarded as inferior by scientists, their employers and policy makers (Abbott et al. 2010). Metrics of scientists' productivity and impact increasingly affect their employment, promotion, pay, research funding and the reputation of themselves and their organization. Publications that are not peer-reviewed have negligible value in such assessments. However, a review of how the European Union may take advantage of the increasing amount of data recognized the rapidly growing volume of data, 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 (Wood et al. 2010). Calls for datasets to be cited in a conventional manner are now widespread (e.g. Sieber and Trumbo 1995, Altman and King 2007, Birney et al. 2009, Costello 2009a, Chavan and Ingwerson 2009, Constable et al. 2010, Cragin et al. 2010, Page 2010, Lawrence et al. 2011, Mons et al. 2011, Tenopir et al. 2011, Whitlock 2011), and an online register that links data sets to Digital Object Identifiers (DOI) has been launched (Anon. 2011b). However, only a small fraction of datasets and online resources have been so cited (Parsons et al. 2010).

While the question of peer-review of data publications has not been considered in any detail, it is now being encouraged (Costello 2009a, Parsons et al. 2010, Lawrence et al. 2011). Chavan and Penev (2011) argued for the creation of 'data papers' which peer-

review the metadata, but only tentatively suggested peer-review of the data themselves, and announced the intention to create a *Biodiversity Data Journal* that only publishes data papers. However, in practice, the journal has found that referees review both the data and metadata (Chavan, pers. obs.). Some digital resources are already subject to independent external peer-review: for example 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 (Lawrence et al. 2011). These examples demonstrate that peer-review of biodiversity data and information are both possible and practical.

Already, datasets published by GBIF must comply with certain standards and undergo technical cleaning and indexing. It would be possible to submit the results of the technical assessment of datasets, the metadata, and supporting statistics, maps and other representations of the data (e.g. tables) to independent experts for peer-review. We propose a process of quality control and fitness for purpose assessment of biodiversity data publications that includes peer-review (Table 1). Data may be visible online at any stage in the process but will not have all the quality indicators. Thus the added quality control options do not impede data publication. Online comment boxes could allow users to comment on the published data, and the authors and editors of datasets to provide subsequent information (e.g. announce new publications that included the data and additional data now published). Peer-review can be preceded by technical checks that the species names are valid, geographic coordinates map correctly, the data appears consistent and complete, and metadata fields are complete. This peer-review may include a list of questions the referees may be asked to answer with respect to their review. For example: Is the description of the dataset complete, clear and adequate to understand the taxonomic, temporal and geographic scope of the data? Does it contain appropriate citation of methods and data analyses (if any)? Is it necessary to have more information on any aspects, for example how or why the data were collected? Can you see specific problems in the way the data are presented that would hamper its re-use? Do the data appear to be of a standard you would expect of a professional in your field? How might the data be used by other researchers? How significant is the dataset in terms of size, scope and uniqueness? Is there a commitment by the dataset authors, editors or their colleagues to respond to enquiries from users and amend the data and metadata as may become desirable?

A concern in adopting peer-review is the availability of willing referees. This is already a problem for science publications. It is remarkable that there are few incentives employed 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, if referee availability becomes a problem in peer-reviewing data, then several options can be explored. For example, public acknowledgement of the referees, invitations to write special articles and/or join editorial boards, payment of honoraria, and/or employment of a few experts as 'in-house' reviewers instead of relying on many unpaid referees. The option of exposing papers to public 'open-review' before and after publication has had limited uptake (e.g. Liu 2007, Gibson 2007), indicating that editors need to be proactive and invite reviewers.

Calls for quality metrics in other fields of science include a proposed reliability score for mass spectrometry data (Gough and Yaffe 2011). Considering the wide range of potential uses of biodiversity data and the principle of fitness for purpose, a 'reliability' index may be difficult to implement. However, measures such as: number of records, species and geographic locations; proportion of species names validated; completeness of all possible data fields; spatial accuracy; completeness of metadata; error rates in taxonomy, geography and other data fields; would be useful to potential users (Costello and Vanden Berghe 2006, Heidorn 2008, Chavan and Ingwersen 2009, Costello 2009a). Examples of such metrics have been demonstrated by SpeciesLink (CRIA 2011).

Section 4: Solutions

Key elements of data publication should follow the established and respected practices of other scientific publications in several regards, including editorial quality control, independent peer-review, citation of the published dataset, and permanent archiving. The metadata descriptors must include authors, their contact details, abstract, keywords and any other necessary information to enable abstracting services to include the publications in their databases (Costello and Vanden Berghe 2006, Altman and King 2007). This more formal approach to data publication will require the more comprehensive metadata, essential for accurate usage of the data (Michener et al. 1997). This should include data provenance, context, precision, and references to papers that used the data, and is likely to require the development of metadata standards and standardized vocabularies. Michener et al. (1997) provided a list of metadata descriptors for ecology data to which

taxonomic metadata (e.g. taxa included in the study) could be easily added. The advantage of the journal model for data publication is that it implies editorial quality control, and it may be more attractive to scientists to publish in, especially if recognized as a 'peer-review' publication.

This process merits consideration of the establishment of one or more online open-access biodiversity data journals, and/or the adoption by existing journals of data publication. There could be several such journals, either competing and/or specializing in different ways, such as photographs to aid identification for individual whales, marine species, species geographic distribution, or population abundance time-series data. They could be linked through portals with other databases, such as GBIF, and to published literature like the Biodiversity Hubs created by the Public Library of Science (PLoS). Already, two sister journals, *ZooKeys* and *PhytoKeys,* submit data to GBIF post-publication of its accompanying paper (Penev et al. 2009). The first such 'data paper' was published in 2011 (Narawade 2011). For a data journal to be widely abstracted and get an Impact Factor, it is important that it be designed as a peer-reviewed scholarly journal; not just for publishing data papers, but for a wider variety of publications of scholarly value so that these will be abstracted and included in journal citation rankings. These may include editorials on topics of interest in the field, 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. All of these papers would cite other papers, including papers that used the data there published, methods used to collect the data, and standards followed by the data. Authors of data papers could receive automated messages when their paper is downloaded and cited, and be contactable by data users such that new collaborative publications may arise.

There are some differences between conventional print journals and data journals that must be addressed. First, a dataset must be published to rigorous domain-specific standards of formatting and structure to enable it to be combined with other datasets, such as by following the GBIF data schema so it can be automatically integrated into GBIF (Wieczorek et al. 2012). Second, datasets will often be supplemented by additional data over time. For example, new versions of datasets may have corrected errors and omissions, and time-series data will add new datasets over time that may be similar to previous datasets in many respects. We agree with others (Klump et al. 2006, Altman and King 2007) that these should be published as new publications because their data will be

unique, the authors and metadata may change, and for time-series data the temporal scope of each publication will be discrete. The publisher may decide to allow datasets to be corrected should errors be found. If minor, this may be noted as a comment to the dataset, but if significant, the publication would be replaced by a new publication. The old version becomes archived in case researchers wish to compare the consequences of different versions on the analyses so that the results of studies using particular versions can be reproduced. As is presently the case in scientific publications, the details of implementation of these options will be at the discretion of the editors of the publication.

Publication costs

Another consideration regarding data publication is who pays for the publication process and long-term maintenance of the data accessibility. This cost may come from readers (users), institutional libraries, authors, or be sponsored by organizations such as government institutions. There are additional costs when data are integrated with other datasets, a service performed by data centres such as GBIF. All data published through GBIF are immediately open-access. While another publisher may provide preferential access to subscribers, as currently happens with scientific journals where they need subscriber funds to pay for services, such restrictions on data availability must be avoided by the scientific community and their funding agencies. The authors of open-access publications are more cited (Gargouri et al. 2010, Harnard 2008), and the open-access business model is more cost-effective for society (SQW Ltd 2004, Houghton et al. 2009). If asked, commercial publishers may make online appendices open-access because such access is likely to attract readers to purchase the accompanying paper.

The support of GBIF, and related databases such as GenBank, by government funding ensures the data are immediately open-access. This means such biodiversity data are accessible to countries, including developing countries where the data may have been collected. Open-access means that third parties are expected to use the data, create new datasets from them, and benefit from them 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 government funds that enabled their publication. Having collective databases like GBIF, OBIS and VertNet is simpler in terms of user access, and the development of standards and analytical tools that facilitate data integration and synthesis. Such combined resources will also be more cost-effective to support, and precedents for tiered financial contributions have already been established (e.g. based on GDP).

Conclusions

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. The formal publication of datasets, including peer-review, is a logical step in scholarly publication, and will assist the evolution of closer integration of publications and databases (Bourne 2005). Indeed, considering that data are the basis of all information and knowledge, it is at least as important that datasets are peer-reviewed as for the papers resulting from their analysis.

Already, papers are linked together by keywords, authors' names, and the papers they cite. The disaggregation of species descriptions in taxonomic papers is now possible, such that the content may be re-compiled in a database to facilitate more advanced content searching and browsing, and the creation of new publications (e.g. Penev et al. 2010). Websites can already generate maps, tables and graphs from data online. One can imagine future tools that will download the data that contributed to a table, graph or map in a paper. However, the implementation of tools usually requires authors to invest time in marking-up their paper which may be (in their opinion) too time consuming.

What should the global biodiversity science community, particularly GBIF, do to facilitate this process? We recommend that GBIF adopt a more conventional 'publication' process involving step by step checks of metadata and data, that enable a quality index to be derived (Table 1). GBIF should have a clearly visible and easily downloaded citation for each dataset, standard agreement with every dataset author or editor, apply a Creative Commons Attribution only licence, visible metrics of data use and error rates, and explore the options to support biodiversity data journals and/or publish 'data papers' itself. Where appropriate, datasets published through the journal would be automatically integrated into the GBIF quality control and indexing workflow. Their publications would be recognized as having passed the highest standard of quality assurance, including peer-review. Such a journal may also publish data outside the scope of the present GBIF database, such as images. This data publication process could start rapidly through the publication of datasets already quality controlled, but not peer-reviewed by GBIF. GBIF should seek the support of ecology and biology journals to direct their authors to publish

datasets through this journal or to GBIF directly or indirectly through one of its participant organizations (i.e. without peer-review) (Moritz et al. 2011). Most journals already have such requirements for molecular data (Costello 2009a) so this principle is established. However, it requires GBIF and similar data centres to be listed as recommended data publishers. Individual scientists, as authors, referees and editors, should support this good practice by appropriately citing data sources, publishing their data, and pressing others to do the same.

While the development of peer-reviewed publication may be novel in the new field of biodiversity informatics, it is not radical. Already several peer-reviewed journals publish primary environmental and biodiversity data, and primary data were published in monographs, cruise reports and appendices to papers in the past. We recognize that this standard of quality assurance may not be practical in all situations. Thus our proposed tiers of quality assurance allow data to be published immediately and thereafter subject to steps of automated, semi-automated, and finally 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 it is to 'publish' and for whom 'making data available' is not a priority.

Acknowledgements:

We thank Rod Page (University of Glasgow) and A. Townsend Petersen (University of Kansas) for helpful discussion and criticism of early versions of this paper, and Hannu Saarenmaa, Peter Desmet, and Tim Hirsch for helpful suggestions.

References:

Abbott A., Cyranoski D, Jones N, Maher B, Schiermeier Q, Van Noorden R. 2010. Do metrics matter? *Nature* 465, 860-862.

Alsheikh-Ali AA, Qureshi W, Al-Mallah MH, Ioannidis JPA (2011). Public availability of published research data in high-impact journals. *PLoS One* 6(9): e24357.

Altman M., King G (2007) A proposed standard for the scholarly citation of quantitative data. *D-Lib magazine* 13 (3/4), doi: 10.1045/march2007-altman.

Anonymous (2011a). Global Invasive Species Database. Accessed at http://www.issg.org/database on 19th December 2011.

Anonymous (2011b). Datacite. http://datacite.org Retrieved 15th March 2011.

Appeltans W, Bouchet P, Boxshall GA, Fauchald K, Gordon DP, Hoeksema BW, Poore GCB, van Soest RWM, Stöhr S, Walter TC, Costello MJ. (eds) (2012). World Register of Marine Species. Accessed at http://www.marinespecies.org on 2012-02-15.

Bénichou L, Martens K, Higley G, Gérard I, Dessein S, Duin D, Costello MJ 2012. *European Journal of Taxonomy*: a public collaborative project in open access scholarly communication. *Scholarly and Research Communication* in press.

Baird R. 2010. Leveraging the fullest potential of scientific collections through digitization. *Biodiversity Informatics*, 7: 130 – 136.

Berents P, Hamer M, Vishwas Chavan 2010. Towards demand-driven publishing: approaches to the prioritization of digitisation of natural history collections data. *Biodiversity Informatics,* 7: 113-119.

Birney E, Hudson TJ, Green ED, Gunter C, Eddy S, Rogers J, Harris JR, Ehrlich DS and 67 others. 2009. Prepublication data sharing. *Nature* **461**, 168-170. doi:10.1038/461168a.

Bisby F.A., Roskov Y.R., Orrell T.M., Nicolson D., Paglinawan L.E., Bailly N., Kirk P.M., Bourgoin T., Baillargeon G., Ouvrard D., eds (2011). Species 2000 & ITIS Catalogue of Life: 2011 Annual Checklist. Digital resource at www.catalogueoflife.org/annual-checklist/2011/. Species 2000: Reading, UK.

Bourne P (2005) Will a biological database be different from a biological journal? *PLoS Comp Biol* 1(3): e34.

Centro de Referência em Informação Ambiental (CRIA). 2011. SpeciesLink. Accessed at http://splink.cria.org.br/dc/index?&system=&setlang=en on 16 December 2011.

Chavan VS, 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

Chavan V, Penev L. 2011. The data paper: a mechanism to incentivise data publishing in biodiversity science. *BMC Bioinformatics* 12 *supl 15): S2, 12pp.

Chavan V, Rane N, Watve A, Ruggiero M. 2005. Resolving taxonomic discrepancies: role of electronic catalogues of known organisms. *Biodiversity Informatics* 2, 70–78.

Chavan, V. S., Gaiji, S., Hahn, A., Sood, R. K., Raymond, M., and N. King. 2010. State-of-the-network 2010: discovery and publishing of the primary biodiversity data through the GBIF network. Copenhagen: Global Biodiversity Information Facility, 36 pp.

Constable, H., Guralnick, R., Wieczorek, J., Spencer, C.L,, Peterson, A. T. 2010. VertNet: a new model for biodiversity data sharing. PLoS Biology, 8(2), e1000309. doi: 10.1371/journal.pbio.1000309.

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

Costello M.J. 2009a. Motivation of online data publication. BioScience 59 (5), 418-427.

Costello M.J. 2009b. Distinguishing marine habitat classification concepts for ecological data management. *Marine Ecology Progress Series* 397, 253-268.

Costello M.J., Appeltans W. 2008. Taxonomic editors plan a World Register of Marine Species (WoRMS). MarBEF Newsletter No. 8, 36-38. ISSN 1649-5519.

Costello, M.J., Vanden Berghe E. 2006 "Ocean Biodiversity Informatics" enabling a new era in marine biology research and management. *Marine Ecology Progress Series* 316, 203–214.

Costello, M. J., Emblow, C. and White R. (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.

Costello M.J., Stocks K., Zhang Y., Grassle J.F., Fautin D.G. 2007. About the Ocean Biogeographic Information System. Retrieved from http://hdl.handle.net/2292/5236

Costello MJ, Harris P, Pearce B, Fauchald K, Fiorentino A, Bourillet J-F, Hamylton S (editors) 2010. A glossary of terminology used in marine biology, ecology, and geology. Version 1.0. Accessed at http://www.marinespecies.org/glossary on 19th December 2011.

Costello M.J., Vanden Berghe E., Browman H.I. (eds) 2006. Introduction: Ocean Biodiversity Informatics. *Marine Ecology Progress Series* 316, 201-202.

Costello MJ, Wilson SP, Houlding B. 2011. Predicting total global species richness using rates of species description and estimates of taxonomic effort. *Systematic Biology* published online 18 August 2011. DOI:10.1093/sysbio/syr080

Cragin MH, Palmer CL, Carlson JR, Witt M. 2010. Data sharing, small science and institutional repositories. *Phil. Trans. R.A* 368, 4023–4038.

Desmet P. 2012. Why we should publish our data under Creative Commons Zero (CC0). Accessed at http://www.canadensys.net/2012/why-we-should-publish-our-data-under-cc0 on 14 April 2012.

Froese, R., Pauly D. (eds) 2011. FishBase. Accessed at www.fishbase.org, version (12/2011) on 19 December 2011.

Gargouri Y, Hajjen C, Larivière V, Gingras Y, Carr L, Brody T, Harnad S (2010) Self-selected or mandated, open access increases citation impact for higher quality research. *PLoS ONE* 5: e13636.

Gaston, K.J., Mound L.A. (1993) Taxonomy, hypothesis testing and the biodiversity crisis. *Proc. R. Soc. Lond.* B 251, 139–142.

Gibson TA. 2007. Post-publication review could aid skills and quality. *Nature* 448, 408.

Global Biodiversity Information Facility. 2009. Participants Report 2009. Global Biodiversity Information Facility, Copenhagen, 71 pp.

Global Biodiversity Information Facility. 2010. Annual Report 2009. Global Biodiversity Information Facility, Copenhagen, 72 pp.

Global Biodiversity Information Facility Secretariat. 2011. Publications that cite GBIF. Accessed http://www.editgrid.com/user/gbif_secretariat/Professional_Publications_that_cite_GBIF on 21 November 2011.

Global Biodiversity Information Facility. 2012. Recommended practices for citation of the data published through the GBIF network. Version 1.0 (authored by Vishwas Chavan). Global Biodiversity Information Facility, Copenhagen, 12 pp. Accessed at http://links.gbif.org/gbif_best_practice_data_citation_en_v1, 22 July 2012.

Gough NR, Yaffe MB. 2011. Focus issue: conquering the data mountain. *Sci. Signal.* 4, 1-3.

Group on Earth Observations (GEO) 2010. Implementing the Data Sharing Principles. GEO News No. 11, http://www.earthobservations.org/art_011_002.shtml.

Guiry, M.D., Guiry, G.M. 2011. *AlgaeBase*. World-wide electronic publication, National University of Ireland, Galway. Accessed at http://www.algaebase.org on 18 December 2011.

Guralnick RP, Hill AW, Lane M. 2007. Towards a collaborative, global infrastructure for biodiversity assessment. *Ecology Letters* 10, 663-672.

Hagedorn G, Mietchen D, Morris RA, Agosti D, Penev L, Berendsohn WG, Hobern D. 2011. Creative Commons licenses and the non-commercial condition: implications for the re-use of biodiversity information. *ZooKeys* 150: 127–149.

Harnad S. 2008. Validating research performance metrics against peer rankings. *Ethics in Sciences and Environmental Politics* 8, 5pp.

Heidorn PB. 2008. Shedding light on the dark data in the long tail of science. *Library Trends* 57 (2), 280–299.

Hill AW, Otegui J, Anino AH, Guralnick RP. 2010. GBIF position paper on future directions and recommendations for enhancing fitness-for use across the GBIF network. Global Biodiversity Information Facility, Copenhagen, 25 pp.

Houghton J, Rasmussen B, Sheehan P, Oppenheim C, Morris A, Creaser C, Greenwood H, Summers M, Gourlay A. 2009. Economic implications of alternative scholarly publishing models: exploring the costs and benefits. Victoria and Loughborough Universities, UK, 256 pp.

Ingwersen P, Chavan V. 2011. Indicators of the Data Usage Index (DUI): an incentive for publishing primary biodiversity data through global information infrastructure. *BMC Informatics* 12 (Suppl. 15): S3.

Jinha A. 2010. Article 50 million: an estimate of the number of scholarly articles in existence *Learned Publishing, 23* (3), 258–263 DOI: 10.1087/20100308

Klump J., Bertelmann R., Brase J., Diepenbroek M., Grobe H., Höck H., Lautenschlager M., Schindler U., Sens I., Wächter J. 2006. Data publication in the open access initiative. *Data Science Journal 5,* 79–83.

Lawrence B, Jones C, Matthews B, Pepler SJ., Callaghan S 2011. Citation and Peer Review of Data: Moving Towards Formal Data Publication. *International Journal of Digital Curation* 6 (2), 4–37.

Liu SV. 2007. Why are people reluctant to join open review? *Nature* 447, 1052.

Los W, Wood J. 2011. Dealing with data: upgrading infrastructure. *Science* 331, 1515–1516.

Merali Z., Giles J. 2005. Databases in peril. *Nature* 435, 1010–1011.

Michener WK, Jones MB. 2011. Ecoinformatics: supporting ecology as a data-intensive science. *Trends in Ecology and Evolution* 1491, 1–9,

Michener WK, Brunt JW, Helly JJ, Kirchner TB, Stafford SG 1997. Nongeospatial metadata for the ecological sciences. *Ecological Applications* 7 (1), 330–342.

Mons, B., Haagen, H. van, Chichester, C., Hoen, P.-B. 'T, Dunnen, J. T. den, Ommen, G. van, et al. (2011). The value of data. *Nature Genetics*, 43(4), 281–3.

Moritz T, Krishnan S, Roberts D, Ingwersen P, Agosti D, Penev L, Cockerill M, Chavan V. 2011. Towards mainstreaming of biodiversity data publishing: recommendations of the GBIF Data Publishing Framework Task Group. *BMC Bioinformatics* 2011, 12 (Suppl.15): S1.

Narwade S, Kalra M, Jagdish R, Varier D, Satpute S, Khan N, Talukdar G, Mathur VB, Vasudevan K, Pundir DS, Chavan V, Sood R (2011) Literature based species occurrence data of birds of North-East India. In: Smith V, Penev L (Eds) e-Infrastructures for data publishing in biodiversity science. *ZooKeys* 150: 407–417.

Page R.D.M. 2006. Taxonomic names, metadata, and the semantic web. *Biodiversity Informatics* 3, 1–15.

Page RDM 2008. Biodiversity informatics: the challenge of linking data and the roleroe of shared identifiers. *Brief. Bioinfor.* 9, 345–354.

Page, R. D. M. (2010). Enhanced display of scientific articles using extended metadata. Web Semantics: Science, Services and Agents on the World Wide Web, 8(2–3), 190–195. doi: 10.1016/j.websem.2010.03.004.

Parr CS, Guralnick R, Cellinese N, Page RDM. 2012. Ecolutionary informatics: unifying knowledge about the diversity of life. *Trends in Ecology and Evolution* 27, 94-103.
Parsons MA, Duerr R, Minster J-B. 2010. Data citation and peer review. *EoS* 91, 297–299.

Penev L, Erwin T, Miller J, Chavan V, Moritz T, Griswold C. 2009. Publication and dissemination of datasets in taxonomy: Zookeys working example. *Zookeys* 11, 1–8.

Penev L, Roberts D, Smith V, Agosti D, Erwin T. (2010) Taxonomy shifts up a gear: new publishing tools to accelerate biodiversity research. *ZooKeys* 50: i–iv.

Peterson AT, Canhos D, Gardenfors U, Scholes RJ, Shirayama Y, Graham MS, Pando F. 2010. *Forward looking report*. Global Biodiversity Information Facility, Copenhagen, 45 pp.

Piwowar HA (2011) Who shares? Who doesn't? Factors associated with openly archiving raw research data. *PLoS ONE* 6(7): e18657.

Piwowar HA, Day RS, Fridsma DB (2007) Sharing detailed research data is associated with increased citation rate. *PLoS ONE* 2(3): e308.

Porter JH, Hanson PC, Lin C-C. Staying afloat in the sensor data deluge. *Trends in Ecology & Evolution, Available online 27 December 2011,*

Reichman OJ, Jones MB, Schildhauer MP. 2011. Challenges and opportunities of Open Data in ecology. *Science* 331, 703–705.

Robert V, Stegehuis G, Stalpers J. 2005. The MycoBank engine and related databases. Accessed at http://www.mycobank.org on 15th January 2012.

Rumble J Jr, Carroll B, Hodge G, Bartolo L. 2005. Developing and using standards for data and information in sciences and technology. In Proceedings of conference '*Ensuring long-term preservation and adding value to scientific and technical data (PV 2005)*', PV 2005, The Royal Society, Edinburgh, 21–23 November 2005. Accessed http://www.ukoln.ac.uk/events/pv-2005 on 20 Novermber 2011, 13 pp.

Saeedi H, Costello MJ. 2012. Aspects of global distribution of six marine bivalve mollusc families. In: da Costa F. (Ed.), *Clam fisheries and aquaculture,*.Nova Science Publishers Inc., New York. In press.

Santos C., Blake J., States D.J. 2005. Supplementary data need to be kept in public repositories. *Nature* 438, 738.

Sieber JE, Trumbo BE. 1995. (Not) giving credit where credits due: citation of data sets. *Sci. Engineer. Ethics* 1, 11–20.

Smit E. 2010. 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. Accessed http://datacite.org/datacite_summer_meeting_2010 on 21 November 2011.

SQW Ltd. 2004. Costs and business models in scientific research publishing. A report commissioned for the Wellcome Trust. The Wellcome Trust, London, 24pp.

Stork, N.E., Grimbacher P.S., Storey R., Oberprieler R.G., Reid C., Slipinski S.A. (2008) What determines whether a species of insect is described? Evidence from a study of tropical forest beetles. *Insect Conserv. Diversity* 1, 114–119.

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., et al. (2011). Data Sharing by Scientists: Practices and Perceptions. PLoS ONE, 6(6), e21101. doi: 10.1371/journal.pone.0021101.

The Gene Ontology Consortium. *2000.* Gene ontology: tool for the unification of biology. Nature Genetics 25(1): 25–9.

Thessen AE, Patterson DJ 2011. Data issues in the life sciences. *ZooKeys* 150, 15–51.

Thiers, B. 2012. Index Herbariorum: A global directory of public herbaria and associated staff. New York Botanical Garden's Virtual Herbarium. Accessed at http://sweetgum.nybg.org/ih on 15th January 2012.

Toronto International Data Release Workshop Authors (2009) Prepublication data sharing. *Nature* 461, 168–170.

Van den Eynden V, Corti L, Woollard M, Bishop L, Horton l. 2011. *Managing and sharing data.* UK Data Archive, University of Essex, Colchester, 36pp.

Vision TJ 2010. Open data and the social contract of scientific publishing. BioScience 60, 330–331.

Ware M, Mabe M. 2009. The STM report: an overview of scientific and scholarly journal publishing. International Association of Scientific, Technical and Medical Publishers, Oxford, 68 pp

Wellcome Trust. 2010. Policy on data management and sharing. Accessed http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTX035043.htm on 20 November 2011.

Whitlock MC. 2011. Data archiving in ecology and evolution: best practices. *Tr. Ecol. Evol.* 26, 61–65.

Wood J, Andersson T, Bachem A, Best C, Genova F, Lopez DR, Los W, Marinucci M, Romary L, Van de Somple H, Vigen J, Wittenburg P, Giaretta D, Hudson RL 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*. A submission to the European Commission. European Union, Brussels, 38pp.

Yesson C, Brewer PW, Sutton T, Caithness N, Pahwa JS, et al (2007) How global is the Global Biodiversity Information Facility? *PLoS ONE* 2(11): e1124.

Box 1. Glossary of terms used in this paper.
Definitions are particular to the present context of biodiversity data publication.

Authorship
The person who wrote a document or other literary work (e.g. scientific paper, poem, essay, lyrics). It may be shared amongst co-authors. It cannot be transferred or sold once created.

Copyright
The automatic legal right of ownership of a document, software, artwork or other work, including datasets and databases. Individual data elements or statements of fact are not copyrightable, but when organised in a database they can be. At first it may be assigned to its creator (e.g. author, artist) or their employer or client who commissioned the work. Subsequently, it may be transferred to others. It may expire after the life of the creator plus 50 or more years depending on the applicable national laws.

(Copy or Technical) Editor
A person who prepares text, images and other materials for publication. They improve the quality of presentation of the material for its intended audience.

Intellectual Property Rights (IPR)
A broad term covering all kinds of creative works, including scientific papers, software, creation of databases, artist and musical works, inventions and patents, trademarks. It recognises the creators and copyright holders (not necessarily the same persons or legal entities), and terms and conditions of how people may use the works (i.e. the licensing arrangements).

License
A legal arrangement whereby the copyright holder permits others to use their creative work or products.

Publication
Material made publicly available through print, digital, sound or other media. This does not mean it is free of copyright, and the material may be provided free, by subscription, or sold.

Public domain
Indicates that creative works and other information is both publicly available (i.e. published) and free of copyright restrictions. Thus it can be used without cost. It could be exploited for scientific, artistic, commercial or other uses. However, to assert authorship without attributing the source would be plagiarism. It is expected good practice to always cite sources, even if in the public domain. Some copyright licenses effectively place materials (e.g. shareware) into the Public Domain.

Ownership
In the present context, this determines who holds the copyright and dictates the licensing arrangements (i.e. terms and conditions of use).

Appendix :

Example of metadata developed within the OBIS community in 2006 to describe datasets that it published. Most components of this user-friendly metadata format could be compiled from existing metadata standards available in International Standards Organisation (ISO), USA Federal Geographic Data Committee (FGDC), and NASA's Global Change Master Directory (GCMD). Standards for Taxonomic and Habitat coverage, Data Collection or Data Source were not available but could be developed to provide a controlled vocabulary. Standard classifications (e.g. Costello 2009b) and definitions of terminology are being developed (Costello et al. 2010).

Order	Metadata term	Example
1.	Database name	BioMar – Ireland: benthic marine species survey
2.	Citation	Picton, B.E., Emblow, C.S., Morrow, C.C., Sides, E.M., Tierney, P., McGrath, D., McGeough, G., McCrea, M., Dinneen, P., Falvey, J., Dempsey, S., Dowse, J. and Costello, M. J. 1999. Marine sites, habitats and species data collected during the BioMar survey of Ireland. In: Picton, B.E. and Costello M. J. (eds), *The BioMar biotope viewer: a guide to marine habitats, fauna and flora in Britain and Ireland,* Environmental Sciences Unit, Trinity College, Dublin. Retrieved [date] from www.iobis.org.
3.	Taxonomic coverage	All species living on, in and near the seabed (benthos), excluding microbia.
4.	Geographic coverage	Republic of Ireland, 200 littoral and 700 sublittoral sites surveyed (each including 1-6 sampling stations).
5.	Temporal coverage	1993–1996 for Republic of Ireland
6.	Habitat coverage	Marine, seashores (littoral), sublittoral seabed
7.	Total distribution records	93,000
8.	Total number of taxa	1,500 species
9.	Collection method	Direct observation on seashores and by scuba divers
10.	Data source	Mainly observations. Reference collection of animals available in the National Museum of Ireland and seaweeds in the Herbarium, Trinity College, University of Dublin.
11.	Abstract	The BioMar project was and remains the largest marine ecological seabed survey of the Republic of Ireland. Standard field survey and data management methods developed by the UK Marine Nature Conservation Review (now part of Joint Nature Conservation Committee) were used. This database was published as a compact disc containing data collected during a national survey that provided the basis for (a) a classification of marine biotopes applicable to the North East Atlantic, and (b) the selection of marine Special Areas of Conservation (Marine Protected Areas).

12.	Scientific Contact	Dr Mark J. Costello m.costello@auckland.ac.nz
13.	Technical contact	
14.	Website	www.ecoserve.ie/biomar
15.	Comment	
16.	Date this form completed	25th September 2005
17.	Publications from this data	McGrath, D., Costello, M.J. and Emblow, C. 2000. The hermit crab, *Diogenes pugilator* (Roux, 1829) in Irish waters. *Biology and Environment: Proceedings of the Royal Irish Academy*, 100B (2), 115–118. Costello M. J., McGrath D. and Emblow C. 1999. A review of the distribution of marine Talitridae (Amphipoda) in Ireland, including the results of a new survey of sandy beaches. In: Schram F. R. and von Vaupel Klein J.C. (ed.)*, Crustaceans and the biodiversity crisis: proceedings of the fourth international crustacean congress, Amsterdam, the Netherlands, July 20-24, 1998.* Brill, Leiden, 473–487. Connor, D.W., Brazier, D.P., Dalkin, M.J., Hill, T.O., Holt, R.H.F., Northen, K.O. and Sanderson, W.G. 1999. Marine Nature Conservation Review: marine biotope classification for Britain and Ireland, Version 97.06. In: Picton, B.E. and Costello M. J. (eds), *The BioMar biotope viewer: a guide to marine habitats, fauna and flora in Britain and Ireland,* Environmental Sciences Unit, Trinity College, Dublin. Costello, M. J. 1998. Experience of the BioMar-LIFE project in the electronic dissemination of marine information. In: Cahill B. (ed.), *Proceedings of the Ocean Data symposium, Dublin 1997.* Irish Marine Data Centre, Marine Institute, Dublin, 8 pp on compact disc. Costello, M.J., Picton B.E., Emblow C., Guiry M., Connor D. 1998. Electronic dissemination of marine biodiversity information collated in databases. In: *Marine Science and Technology Programme experiences in project data management*, M. Bohle-Carbonell (ed.), European Commission, Luxembourg, 73–84 pp. Costello M.J., Emblow C.S. and Picton B.E. 1996. Long term trends in the discovery of marine species new to science in Britain and Ireland. *Journal of the marine biological association of the United Kingdom* 76, 255–257. Costello M.J. 1995. The BioMar (Life) project: developing a system for the collection, storage, and dissemination of marine data for coastal management. In: Hiscock K. (ed.) *Classification of benthic marine biotopes of the north-east Atlantic. Proceedings of a BioMar - Life workshop held in Cambridge 16–18 November 1994.* Joint Nature Conservation Committee, Peterborough, 9–17. Costello M.J. 1993. Development of the BioMar database, and its contribution to nature conservation management in the Irish Sea. In: *Marine and Coastal databases*, Irish Sea Forum Seminar Report No.3, Liverpool University Press, Liverpool, 72–79.