

Motivating Online Publication of Data

MARK J. COSTELLO

Despite policies and calls for scientists to make data available, this is not happening for most environmental- and biodiversity-related data because scientists' concerns about these efforts have not been answered and initiatives to motivate scientists to comply have been inadequate. Many of the issues regarding data availability can be addressed if the principles of "publication" rather than "sharing" are applied. However, online data publication systems also need to develop mechanisms for data citation and indices of data access comparable to those for citation systems in print journals.

Keywords: data availability, online publication, environment, citation indices, biodiversity informatics

Scientists advance knowledge gained from empirical and modeled data and observations. It follows that scientists who do not publish or release their data are compromising scientific development and, arguably, leaving their work unfinished. Considering that science is based on observations, it is astonishing that the publication of primary data is not a universal and mandatory part of science. The reasonable expectation of society that science will make data available for further research—especially if the research that produced the data was publicly funded—is supported by a wide range of international and national policies, and in principle by the science community and publishers. If data are not made publicly available or lodged in a permanent archive like any unpublished research, they are likely to be lost over time (Heidorn 2008).

The availability of data from local to global scales is critical for dealing with current issues affecting society, such as climate change, public health, and biodiversity loss. Society expects that scientists will make their data available because most data are paid for directly (i.e., government funded) or indirectly (e.g., university salaries) by public funds, or are collected for the public good (e.g., public health, product safety, environmental monitoring data). This interest of society is demonstrated by the *Guardian* newspaper campaign for the release of government-funded geographic data in Britain (www.guardian.co.uk/technology/freeourdata and www.freeourdata.org.uk/blog) and by the emergence of organizations such as the Open Knowledge Foundation (www.okfn.org).

Many international, intergovernmental, and funding agencies have policies calling on member countries or grant recipients to make data available (Edwards et al. 2000, Arzberger et al. 2004a, 2004b, Costello et al. 2008); among these agencies are the Organization for Economic Cooperation and Development, the International Council for Science (ICSU), the Intergovernmental Oceanographic Commission (IOC) of the United Nations Educational, Scientific and Cultural Organization, the Global Biodiversity Information Facility (GBIF), the European Research Council, UK Research Councils, and the US National Science Foundation and National Institutes of Health, and even some treaties (such as the Antarctic Treaty System; www.scar.org/treaty). Some journals, including *Science* and *Nature*, explicitly expect data to be made publicly accessible, and they list suitable repositories for certain types of data. The Association of Learned and Professional Society Publishers and the International Association of Scientific, Technical and Medical Publishers (2006) recommend public access to data that support publications.

A comparison of national policies regarding the availability of government data showed that open access conferred significant economic benefits by stimulating entrepreneurial use of the data by commercial companies (McMahon 1996, Weiss 2002). In contrast, restrictive data-release policies and fees for data use (which provide negligible financial return) discouraged innovation and development of data products. However, despite the recognized societal benefits, most primary data remain unavailable. For example, a recent review of national ocean data centers, part of a 30-year-old network

established by the IOC, found that the centers generally had less than half the data they should have for each country, and many countries still lack such data centers (Kohnke et al. 2005). More than 70 percent of the organizations publishing data through GBIF and the Ocean Biogeographic Information System (OBIS) are from government organizations (including museums), and less than 20 percent are from universities and individual scientists, which may reflect the greater influence of government and international policies on the former group of organizations. The policies and calls for data sharing have not been sufficient to make data sharing the normal practice throughout science.

More than 70 countries and 50 other organizations make up the Group on Earth Observations (GEO; www.earthobservations.org), which aims to establish a Global Earth Observation System of Systems (GEOSS) by 2015. The GEOSS will cover all observation data, from climate to biodiversity, including those recorded by satellites, buoys, *in situ* sampling, and observations; GEOSS data will be available and integrated through a common portal (www.geoportal.org). However, this system will be successful only if all data are readily available: historic data will have to be digitized and the world scientific community will have to contribute data through common standards, protocols, and open-access agreements (Scholes et al. 2008). Although it has recognized this problem, GEO has not yet proposed a solution.

When new species, proteins, gene sequences, microarrays, cell lines, and bacterial strains are described, the scientific community expects that type specimens will be deposited in suitable collections (e.g., museums, herbaria) and molecular data will be deposited in specialist data centers (e.g., Protein Data Bank in the United States, Cambridge Crystallography Data Centre in the United Kingdom, GenBank), and most journal editors make such action a prerequisite for publication of a print paper. Howe and colleagues (2008) list 21 molecular biology databases on model organisms and 3 additional ones with data on numerous species. GenBank comprises mirror sites at the European Molecular Biology Laboratory, the National Center for Biotechnology Information in the United States, and the DNA Databank of Japan. Each database is government funded and all the data are freely accessible online. Thus, making data publicly available is already part of the culture in some sciences, such as physics (e.g., the *arXiv.org* preprint series), astronomy, climatology, and molecular biology (RIN 2008). However, even in such well-established fields as bioinformatics, in which one can get a degree and become a “biocurator,” incentives such as improving the citability of contributions have been called for (Howe et al. 2008).

Very large databases are curated by professional data managers, because the highly standardized data in them (collected automatically by sensors on satellites, buoys, or other platforms) demand it (Heidorn 2008). A similar amount of more diverse data spread through many small data sets and individual scientists is not being professionally curated (Heidorn 2008), yet the size of a data set is not necessarily an

indicator of the data’s value to science now or in the future. If some of these small data sets could be standardized, they could be published through facilities such as GenBank and GBIF. The development of more standards for publication of different data types is thus to be encouraged.

Although governments, funding agencies, and the scientific community appreciate the benefit of making data publicly available, individual scientists may not find the benefits quite as evident. This is because individual scientists’ concerns about making data openly available and introducing measures to motivate online publication have not been addressed (Klump et al. 2006, Parr 2006, Blagoderov et al. 2008, Heidorn 2008, RIN 2008). The main obstacle to making more primary scientific data available is not policy or money but misunderstandings and inertia within parts of the scientific community. In this article, I seek to answer the responses I have heard repeatedly from scientists when asked why they do not publish their data online. Their reservations must be addressed to change scientists’ behavior from data hoarding (and occasional data sharing) to online data publication.

Some benefits of data publication

Online data publication will boost scientists’ recognition, generate invitations to meetings, present consulting and collaboration opportunities, and increase citation rates because their productivity will be more visible (box 1; Froese et al. 2004, Eysenbach 2006, RIN 2008). Compared with publishing data in print media or archiving it in libraries, publishing data online is less expensive and it exposes the author’s work to a far wider audience. Making data available online maximizes the potential return on the investment in research, and those data can be repatriated to the countries from which they may have been collected by foreign scientists. The cost of saving and reusing data published online is also likely to be lower than the cost of collecting them again (Heidorn 2008). Without the ability to reanalyze the original data from which a scientific conclusion was reached, the conclusion cannot be independently tested (Casey and Blackburn 2006), and some data cannot be replicated because of unique combinations of environmental conditions (Heidorn 2008). Furthermore, making data availability mandatory may help discourage or expose scientific misconduct (Klump et al. 2006). Concern over the modification of images published in science journals has led to recommendations that the primary data and images be made available (Couzin 2006). Such calls would be unnecessary if primary data, whether alphanumeric, sound, or images, were automatically available on the Internet by the time of print publication.

Data publication can also bring benefits at a corporate level. If an organization is required to provide data to the public upon request, making data publication a routine practice can eliminate the tedium of attending to individual data requests piecemeal. Efforts to disseminate data sets through license agreements can also be time consuming, and because user needs vary, it can be difficult to standardize these agreements online without raising questions about liability

should the data be incorrectly used (Freeman et al. 1998). Instead of licenses, “publication” is simpler conceptually and practically, and responsibility for use of the data more clearly lies with the reader.

In contrast to interpretations and opinions derived from data, the value of primary environmental and ecological data grows in time as they become harder to replace. Such data are inevitably a sample of what could be collected at different spatial scales and over time. Comparing new data with other data collected in the same or different places and times may reveal previously unknown patterns over larger areas and timescales. This immediate added value can be further multiplied by the opportunities provided for unforeseen uses and benefits, as found for genomic and proteomic data (Smalheiser 2002).

Why more data are not publicly available

In box 2 are a dozen reasons scientists gave me for not making their data publicly available online. They have been compiled from numerous meetings with researchers over

Box 1. The benefits of online data publication to the participants in research.

Individual scientists as a data creator

- Additional publications
- Greater citation rate
- Wider recognition among peers
- Invitations to meetings
- Invitations to collaborate
- Invitations to provide consultancy

Individual scientists as a researcher and author

- Creators of data are known from citation and so are contactable for more information
- Citation of data sources adds authority that indicates their quality

Editors, peer reviewers

- Independent verification and qualification of research findings is possible

Publishers

- Citation of data publications is likely to increase citations of related research papers

Data centers

- Increased value and role in science, and hence support from the scientific community and funding agencies

Scientific community

- Data can be reused for similar and new purposes
- Data can be integrated with other data to create new data resources

Funding agencies

- Better financial return from research investment as a data can be used again

Governments

- Data are easily accessible to government science advisors

Society

- Better science

the past decade. Although these statements do not constitute a quantitative survey of the community, they are considered representative. Indeed, some of these reasons were also reported in a survey of ocean data centers (Kohnke et al. 2005), and all arose in a survey of UK researchers (RIN 2008) published while this article was under review. The relative frequency or importance of the reasons is not considered, because they have a common solution—namely, to follow the practice of publication rather than data sharing.

There may be valid reasons for not publishing in any form, such as significant errors in the data, protection of individual privacy with medical or survey data, threats from over-exploitation of species or resources, national security concerns, or matters subject to legal action. However, these concerns can be overcome by delaying publication for appropriate periods (Glover et al. 2006), generalizing the data in some way (e.g., giving only a region for the location of a rare species), or not publishing all of the data (e.g., excluding data allowing personal identification).

Too often, scientists release or make data available with conditions that restrict their use or distribution, and thereby create obstacles to their use. Such conditions may be the requirement that the data not be used without the author's permission, that they not be used for commercial purposes, and that any use requires coauthorship on any publications that arise from the data. The same scientists make no such conditions when they publish in print media, and they usually sign away copyright to the publisher and pay the publisher page charges for this service. In contrast, organizations involved in online data publication let copyright remain with the data providers, and to date they have not charged publication fees. The quasi-release of data by attaching conditions to their use is unnecessarily cumbersome, contrary to the scientific publication process, a disincentive to others to explore their potential, and often impractical to enforce.

The term “commercial” is rarely defined and is subject to different interpretations. A developing country may argue that any knowledge gained by scientists in a developed country profits their “knowledge economy” and may result in direct or indirect commercial benefits. A scientist may profit personally by gaining professional promotion or obtaining research funding as a consequence of a paper published from the data. Furthermore, if a person or organization should use the data to produce new knowledge or products that can be sold, they should be compensated for creating added value. It may be difficult to distinguish what is commercial on the basis of a scientist's employer. Some research institutes are legally commercial companies (e.g., the National Institute of Water and Atmospheric Research in New Zealand), and government agencies and universities often do contract work for commercial companies.

As is the case for print media, there should be no discrimination as to who should have access to data published online. In turn, the requirement for data publication should apply to all instances in which the data served as the basis for published papers, regardless of who funded or conducted the study. For

Box 2. Responses to reasons scientists have made for not making data available.

1. *People will copy my work from the Web and plagiarize it.*

The unconventional nature and ease of copying electronic media make some scientists uncomfortable about online publications. Most print-based science journals are now available online, so the potential of copying is already present. By publishing the data, the source will be widely apparent to the scientific community, and thus plagiarism is more likely to be discovered.

2. *Where can one publish data? Journals will not publish primary raw data.*

An increasing number of (and most of the top) journals publish online appendices that can include primary data. In addition, data centers will archive and make publicly available an increasing range of biological and environmental data.

3. *It is my data—why should I make it available?*

In most cases, the scientist is the custodian of data either owned by their employer organization, or, if the data collection was paid for with public funds, then the data should belong to society. Unfortunately, some scientists believe they personally own data even though their salaries and research are funded by their employer or government sources. Most scientists are stewards of public data, and have a responsibility for its dissemination.

4. *The data I used were not my own and I did not get permission to publish them.*

If the data owner provided the data to the scientist, it is possible that they would provide permission to publish it. The question is, then, why did not, or will not, the scientist ask for such permission? Perhaps because they see no personal benefit in that, it will take time from more profitable activities, or they feel no ethical obligation to make data access convenient to others.

5. *If I release data, then I may be scooped by somebody else producing papers from them. I have not finished analyzing the data and I may do further analysis on them.*

In many cases it may be valid to delay publication of data until they have been analyzed and the synthesis published. However, the data should be published before or upon release of the first print publication so their location can be cited. If no publications transpire after 12 months (the convention in astronomy; RIN 2008), then that likelihood decreases in time and it may benefit the scientist more to publish the data so as to obtain some credit for his or her work.

6. *Somebody will use my data and benefit from such use, and worse still, they may be a commercial organization or consultant.*

Scientists want people to benefit from the results of their research when published in print media, and should have the same hopes when publishing data. As with print media, once published, authors have no control over who may use their findings for whatever purpose, and it may be a waste of energy to attempt to do so. Generally, scientists welcome all use and citation of their research in print media, even by competitors and the popular press, so why should they deny such use of data?

7. *The publisher may profit.*

Scientists commonly sign away copyright to for-profit print media publishers. In contrast, most data publication is by government-funded or not-for-profit organizations. If they benefit from the data publication, they will probably reinvest the funds in the same enterprise.

8. *I fear that the data will be used for an incorrect purpose.*

Data sets should be published only with sufficient “metadata” or documentation that describes how they were collected and their limitations. Should this be provided, then, as with print publications, it is up to the users of the data to be sufficiently competent to use and interpret them.

9. *I do not have the skills to publish data on the Internet.*

Anybody who has the skills to manage their data in tables and spreadsheets has sufficient skill to provide such data in a standard format to organizations that will publish data on the Internet.

10. *Intellectual property rights related to data and databases differ between countries.*

Another concern is loss of intellectual property rights (IPR), ownership, authorship, or control of data by making them available. This encourages comparisons between countries of different IPR laws related to data, databases, and publications. The resulting complexity of issues further discourages data sharing. However, this seems unnecessary if data publication is considered in the same manner as a print publication. Such issues do not significantly limit publication in print media, and neither should they on the Internet. The IPR model of some data publication organizations that regularly crawl data sets distributed at many sources is that copyright and ownership stay with the data source. Thus online publication can give data custodians greater control over data publication than can conventional print media.

11. *I will not get due recognition for creating the data.*

Publication is the most certain way that scientists can get public credit for their ideas and work. Similarly, the best way to ensure recognition for collecting useful data is by publishing them.

Science will benefit most if data are published under “open content” policy as described in the Creative Commons licensing agreements. A common requirement of users is attribution of the source, including its authors or editors. If data users clearly cite their sources, as they would for print publications, it adds credibility to the data so used. If authors do not cite data sources, then, as in print media, they may be guilty of plagiarism.

12. *Other reasons.*

As with the print publication process, data publication can expose problems with data and improve its quality. Sometimes data is organized in an idiosyncratic manner that would make it difficult for anybody else to analyze. Rows and columns in tables may not be adequately labeled, consistently formatted, or sufficiently described for other users. Authors may fear that their selective use of data, or possible errors in analysis, may be revealed by data publication. However, exposure of data to independent analysis would benefit science by either providing independent support or further refinement of the originator’s conclusions, or alternative interpretations.

example, if a company wishes to publish a paper with graphs and statistics demonstrating the safety and efficacy of its new method or product, it should also be required to publish the data on which the results were founded.

How to motivate online data publication

The primary motivations for individual scientists to publish in print are to demonstrate their contribution to science, and the consequent peer-recognition that influences one's reputation and employment opportunities, promotion at work, and ability to win further research funding. Other factors may also exist, such as personal satisfaction in completing a study and enthusiasm about communicating findings and opinions to society. These motivating factors should also be brought to bear on data publication.

One common metric of peer-recognition is citation of papers. Citation also shows who is responsible for the information cited and provides its authority, a key aspect of quality assessment. There is a concern that data sets will not be cited in the same way that print publications should be when they are the source of information. This concern is justified, as most online databases do not provide a citation for each data set in a manner similar to that of print media, and data users tend to cite the Web site URL (Uniform Resource Locator) where the data set is found rather than the actual data set and its authors or editors, regardless of whether this information is available. Such incorrect citation is equivalent to authors' citing a journal rather than the papers published in that journal.

There is a precedent for this failure to cite the original source. The publications that describe new species are rarely cited when the species are mentioned in subsequent studies. Indeed, even the practice of citing identification guides and sources of species nomenclature in scientific publications seems to have waned (Agnarsson and Kunter 2007). If they were cited, taxonomic papers, revisions, and identification guides would be among the most highly cited publications, and they would have very long citation lives (Minelli 2003). To better recognize the contributions of taxonomists to science, different metrics are required, such as how often a species name is used both overall and in particular fields of study, such as agriculture or genetics.

Data are diverse in origin and format. They may (a) be biological, chemical or physical; (b) constitute environmental or physiological measurements by instruments, experimental results, or observations of species, animal behavior, and phenomena; (c) be derived or modeled from primary data; and (d) take the form of numerical, text, sound, or image files. Their value may be in being a reference or baseline, or in their potential for combination with other data to create new data sets (RIN 2008).

Data linked to species names can be published in GBIF, OBIS, Scratchpads (Roberts et al. 2008), and related systems that integrate standardized species data (e.g., Mayo et al. 2008). Physiochemical ocean data (including primary and model data) can be archived in IOC's network of ocean data centers, which increasingly make these data available online.

The ICSU World Data Centers can accept a wide range of biological and environmental data. The old, pre-Internet model of data centers as archives of data is changing to one that provides an editorial service of quality control—which adds value—and online data publication. Users should have the opportunity to examine the original data, and to easily combine the data with other data.

Tracking data

The increased interoperability and linking between online resources can mean that data may be visible from several locations. The original source of data should be the basis for citation. To facilitate citation, data centers should track data access using automated tools and should display the results on their Web site (Costello and Vanden Berghe 2006, Blagoderov et al. 2008, Heidorn 2008, Roberts et al. 2008); that is, an index should be maintained that tracks data viewed, searched, downloaded, linked to, or cited. Thus, providers can refer to the Web site to see how often their data set was accessed. Authors of publications that use such data should cite the data sets in their list of references, as they do for print media.

There are several methods to track the origin of data. The unique ISBN (International Standard Book Number) and ISSN (International Standard Serial Number) assigned to a printed publication can be used to track and locate the product in bookshops and libraries. However, ISBNs and ISSNs are not assigned to individual articles within a journal. Because the URLs used for Web addresses change over time, registration systems for unique and persistent identifiers of items published on the Internet are being developed (Beit-Arie et al. 2001). A centralized registry now provides and administers a unique identifier for geoscience samples, the International Geo Sample Number, or IGSN (www.geosamples.org). The Handle System (www.handle.net) codes resources—whether journal articles or metadata—so that if their location changes, users can use these codes to find the items at the new URL. A development from the Handle System is the Digital Object Identifier (DOI), which is now widely used by journals and abstracting services to identify papers and their appendices published online; DOIs link to a full citation (i.e., author, title, etc.), and although the DOI is unique to the publication, more than one DOI representing the same item or object may arise (e.g., as would happen if different indexing services assign DOIs to the same publication). The PANGAEA information system at the World Data Center in Germany uses DOIs for primary data sets (Klump et al. 2006); corrected or updated versions of a data set receive a new DOI.

Automated methods to assign globally unique Life Science Identifiers (LSIDs) have been demonstrated for species names (Page 2006). Resolvable LSIDs for tracking species names have been implemented for the Catalogue of Life, Index Fungorum, and ZooBank using ontology standards developed by the Taxonomic Data Working Group. The LSIDs could, in principle, be used for data sets (Orme et al.

2008), but some organization would have to assume responsibility for creating and maintaining the registration system to ensure automated resolution of the identifying numbers, or the community could adopt one of the existing systems, such as the DOI.

For online published data to be cited and abstracted as scientific print papers are, the data set would need to clearly display the following information: author or editor, author's address, the data set's informative and unique title, abstract, keywords, a list of publications related to the data (e.g., publications describing methods or analyses derived from the data), and the name of the online publication Web site (Testa 2004). The data publisher should demonstrate scientific editorial standards, including transparency of the editorial process, names and addresses of editorial board members, quality control procedures, a peer-review system, and a list of data sets published and details about them; the online data publication should be open to international contributions (i.e., it should not be an in-house publication). The publisher should archive the data publication indefinitely at a publicly accessible location, such that future researchers can access the data that were used by others. Online data publications can conform to most of the typical publication standards for print journals, but there are important differences. Notably, in contrast to print papers, a data set published online may be corrected or enlarged over time and thus have several versions, and its size is better measured in data units or bytes than in pages. The more dynamic nature of electronic publications allows them to improve in quality and quantity over time.

The future for online data publication

Printing machines were invented more than 500 years ago. Anyone with the means could print anything they wished. In time, editorial and peer-review systems for scholarly publications came into being and quality improved. Similarly—but within the past 20 years—the Internet has allowed many people to publish whatever they wish on the World Wide Web. Editorial and peer-review systems are now evolving, and they will set a quality mark for online publications. Already, most print-based science journals publish online. Scholarly online data publication should include editorial oversight, standard formats and vocabularies, quality control checks, the ability to correct data found to be in error, quality indicators, and peer review (before or after publication). As with print media, the online data publication process must ensure that data survive and are accessible, that their integrity is maintained, and, critically, that they are citable.

Increasingly, environmental data collected by instruments on, for example, monitoring stations, satellites, buoys, and research ships can be immediately and automatically uploaded to a data center (e.g., Glover et al. 2006; see also National Ecological Observatory Network, www.neoninc.org/about-neon/overview.html). This ensures that the data are backed up, timely, and ready for use immediately (where appropriate, as with weather data) or for release after a certain period. Thus,

where possible, the automated publication of data immediately upon collection is to be encouraged.

Journals that specialize in data publication are emerging, such as *Acta Crystallographica E* in chemistry, *Data Briefs* of the electronic earth science journals *Geochemistry*, *Geophysics*, *Geosystems (G⁵)* and *Earth System Science Data*, and *Ecological Archives* in ecology. Nonnumerical data, such as text and images describing species, can be published online using Scratchpads (<http://scratchpads.eu/>; Roberts et al. 2008). Ideally, as in these examples, data should be open access and in standard formats if these exist for the type of data published. Such journals publish data sets with a citation, abstract, and associated information, as papers are published. This information gives clear credit to the data creators and makes it possible to search for the data sets through bibliographic databases. Because users with this information are likely to cite the online data sets just as they do print papers, the data sets will enter the system of citation statistics. There is no reason in principle that data centers could not similarly provide conventional citations. Indeed, Scratchpads and OBIS do so, and GBIF is considering it.

Copyright issues are less likely to compromise data publication than they are in the print media, because facts, names, and short statements are not copyrightable, although some names and phrases may be trademarked. Thus, information is routinely extracted from the literature without infringing copyright, and may then be compiled into databases through manual or automated means. For example, descriptions of species are not “literary and artistic works” in the sense of copyright legislation, because they are formulated in a standardized language along standardized criteria. They can therefore be excerpted without infringing copyright and republished (Agosti and Egloff 2009). They may then be reagggregated into databases to provide guides to species identification and facilitate online taxonomic collaboration (Mayo et al. 2008).

Data centers usually add significant value to data sets through quality control procedures, ensuring adequate metadata, aggregating data from different sources, and providing online tools to explore, visualize (e.g., maps, graphs), and download the data in formats suitable for further research. Libraries may also archive data in print (and perhaps electronic) form, and some institutions now provide archival services for data. However, data deposited in libraries or institutional archives (or repositories) and published as appendices to journal articles do not get the same editorial quality control and peer-review attention as either journal articles or data lodged in special data centers. In other words, data centers can provide quality control as publishers do for print media and archiving as provided by libraries, and they add value through data integration, indexing, exploration, and visualization services. Preferably, data holders will publish not on their own Web site—where long-term maintenance can be an issue—but in international specialized data centers (e.g., GenBank, GBIF) that will maximize data availability and give it added value. This is the policy of the American Geophysical Union for its journals, and journals such as

Proceedings of the Royal Society and *Nature*. The latter requires data to be sent to the journal for publication; data “cannot be hosted solely on the authors’ own websites.”

When data sets are published, they may be described using a standard set of information fields such as the “Dublin Core” metadata (and by an extension of it called “Darwin Core” if it includes biological species information). Increasingly, authors are required to enter their names, contact details, keywords, and abstracts into Web-based forms when submitting papers for publication. One can envisage this metadata being extended to provide standard descriptions of online data sets and key terms (e.g., name of a species newly described), which can be forwarded to abstracting services and other databases. This metadata is invaluable for allowing people to discover data sets that may be useful to them, but the metadata may not be sufficient to enable them to use the data. Procedures for publication of “use metadata” were recently described by publishers of geochemical journals at a meeting of Editors Roundtable on 16 July 2008 at the Goldschmidt Conference in Vancouver, Canada.

Data should not only be published, they should be published in a way that facilitates integration with other data, that is, in a standard, atomized format on the World Wide Web. Although not all data are easily integrated with other data sets, such as laboratory experiments, the low cost of online publication means that these data can still be published in a nonintegrated way (e.g., as an online appendix that future integration services may use). Where suitable online publishers do not exist for data, authors may publish them in data centers and, less ideally, as online appendices. The latter are generally not as useful as data centers because they lack standards for file formats, data organization, and metadata (Santos et al. 2005).

New data integration services are emerging, such as for geological maps (www.onegeology.org). In addition to the physical and geochemical sciences, scientists with interests in evolutionary, ocean, and biodiversity data have initiatives under way to further the publication of data of interest to them: the (a) National Evolutionary Synthesis Center in the United States, (b) Scientific Committee on Ocean Research and IOC’s International Oceanographic Data and Information Exchange (Costello et al. 2008), and (c) GBIF, respectively.

As is the case in print media, researchers and journal editors need to judge which data merit publication. These decisions could be guided by criteria such as whether specialist publishers exist for the data in question, whether others have published similar data, and whether the data are needed to enable independent reproduction of study findings.

The only valid reasons for scientists not to publish their data online are the same as for not publishing in print media—namely, the data are of such poor quality that they could have no useful purpose, scientists lack the competence or time-management skills required to prepare data for publication, or publication is not a priority in the scientists’ work or career. Thus, those who fail to publish data online should be viewed in a similar light as those who do not publish in

print media. Withholding data after they have been analyzed and a study has already been published, with the intention of professionally profiting further, raises ethical concerns about whether the scientist is really motivated to advance science.

The next steps for data publication

The well-established and successful contemporary model of publishing scientific findings should be complemented by a system of data publication, ideally through data centers, in a way that enables the scientific creators to be credited and cited (figure 1). Greater accessibility and reuse of data will provide additional resources for research, and hence greater benefits to science and society. However, benefits to individual scientists will be fully realized only if the data are published formally and cited by users (box 1). The following actions are critical for full data availability:

- Before data collection, principal investigators must plan for data publication so the preparation of the data for publication is simplified and low cost.
- Scientists involved in the peer-review process should ask that, where appropriate, the data on which studies were based be publicly accessible (without preconditions) so they may be subject to independent analysis and their findings reproduced.
- Journal editors should require authors to publish their data online in standard formats, and, where available, through data centers that offer integration and archival services.
- Online data centers should publish clear, standard citations for data sets; track data-set access; and develop editorial processes to maximize data quality, data integration, accountability, visibility, and usability.
- Authors must cite online data sources as they would print publications.
- Citation services must include online data publications in their metrics.
- Employers of scientists must recognize the efforts of those who publish their data online as they do those who publish in print media, question why scientists have not published their data online, and include data publication as a measure of productivity and performance.
- Funders of research must (a) ensure that research proposals have a data management plan and an appropriate budget for data publication, (b) contractually require data publication upon completion of a project, and (c) withhold further funding from contractors who have failed to fulfill this requirement.

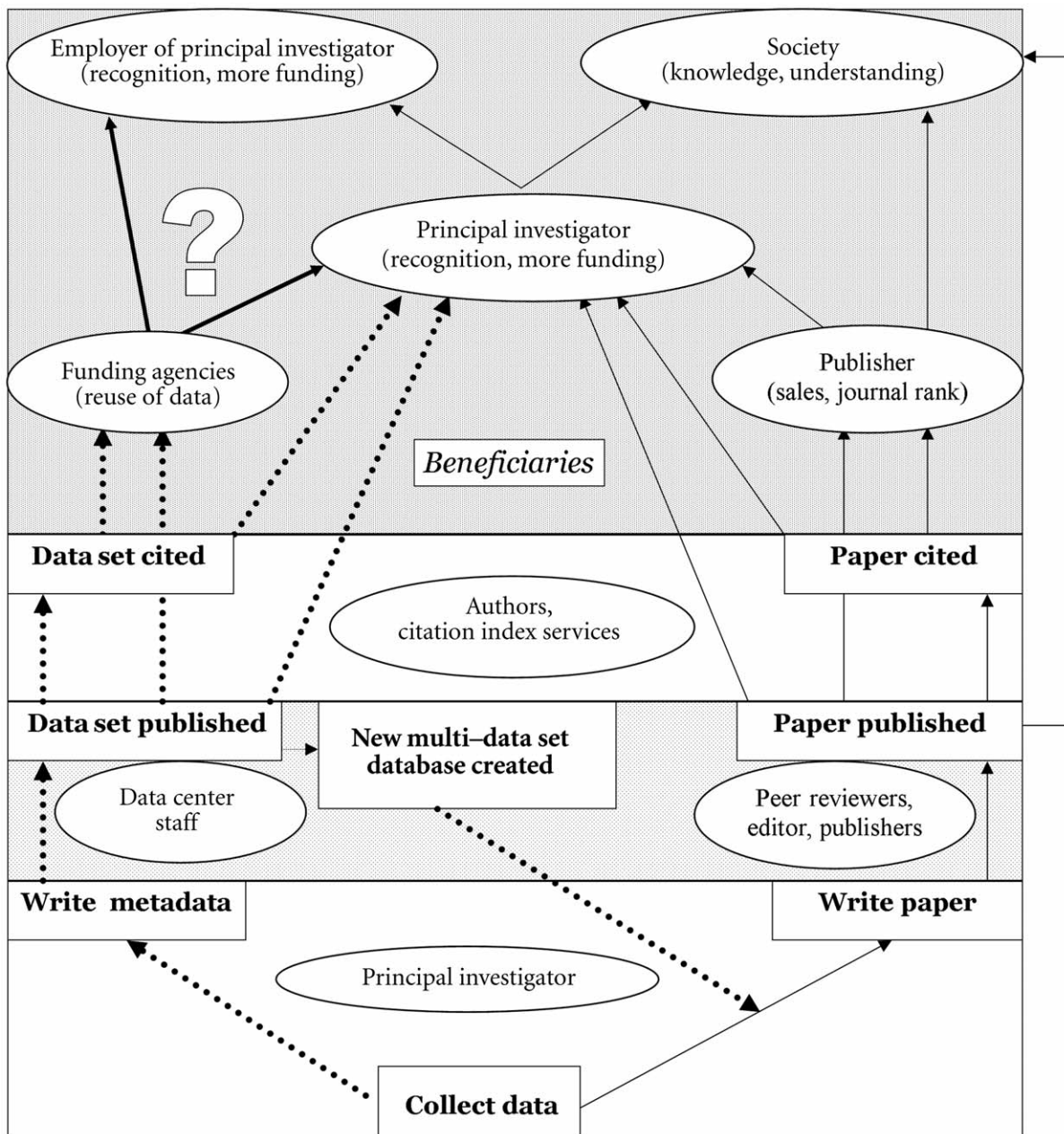


Figure 1. Illustration of the present process of publication (solid lines) and that proposed for data publication (dotted lines). Arrows flow from data collection to the benefits to principal investigators (PIs), their employers, publishers, and society (actors). Key events (actions) are shown in boldface in boxes, and actors in oval shapes. Publication in data centers can provide new resources for research and result in new publications. Should funding agencies insist on data publication, then employers and PIs will benefit from eligibility for additional funding. Should the published data be citable, then recognition of the PIs, and consequently their employers, will be another benefit (box 1). Society benefits from access to the resultant publications, data analyses and interpretation, and scientific advice.

- Governments must financially support online data publication centers.

The main problem in data availability is not a lack of policy, technology, financial resources, or publication outlets, although data centers do need financial support (Merali and Giles 2005). Rather, it is that the science reward system has not kept pace with the new opportunities provided by the Internet, and does not sufficiently recognize online data publication. A change in science culture as a result of the Internet is under way (Kinne 1999, Costello and Vanden Berghe 2006), and we must adapt approaches to scholarly publication accordingly. A confluence of the availability of open-access online resources with the quality control systems that professional editorial processes bring, may be the optimal way forward.

Acknowledgments

The ideas and arguments in this article have benefited from discussions with many other scientists in workshops, meetings and e-conferences over the years, including Donat Agosti, Christos Arvantidis, Scott Baker, Bill Ballantine, Frank Bisby, Robert Branton, Cliff Cunningham, Yde de Jong, Michael Diepenbroek, Jim Edwards, Willi Egloff, Chris Emblow, Mari-Claire Fabri, Daphne Fautin, Rainer Froese, J. Frederick Grassle, Hannes Grobe, Michael Guiry, Peter Herman, Don Hobern, Brewster Kahle, Meredith Lane, Wouter Los, Allan Rodrigo, Karen Stocks, Marc Taconet, Edward Vanden Berghe, Lawrence Way, Peter Wiebe, Cisco Werner, and Yunqing Zhang, and participants in meetings of the European Register of Marine Species, Ocean Biogeographic Information System (OBIS), Census of Marine Life, Marine Biodiversity and Ecosystem Function research network (MarBEF), International Ocean Data Information and Exchange, Coordinated Research on the North Atlantic Project, Scientific Committee on Ocean Research, Diversitas, and the Global Biodiversity Information Facility. I thank Cynthia Parr, Dave Roberts, and the anonymous reviewers of this manuscript for their suggestions, which improved this article. This article contributes to OBIS and to the European Union projects Pan-European Species-directories Infrastructure and the European Distributed Institute in Taxonomy; it is MarBEF publication number MPS-08045.

References cited

Agnarsson I, Kuntner M. 2007. Taxonomy in a changing world: Seeking solutions for a science in crisis. *Systematic Biology* 56: 531–539.

Agosti D, Egloff W. 2009. Taxonomic information exchange and copyright: The Plazi approach. *BMC Research Notes* 2009, 2: 53. doi:10.1186/1756-0500-2-53

Arzberger P, Schroeder P, Beaulieu A, Bowker G, Casey K, Laaksonen L, Moorman D, Uhlir P, Wouters P. 2004a. Promoting access to public research data for scientific, economic and social development. *Data Science Journal* 3: 135–152.

———. 2004b. An international framework to promote access to data. *Science* 303: 1777–1778.

Association of Learned and Professional Society Publishers and the International Association of Scientific, Technical and Medical Publishers. 2006. Databases, data sets, and data accessibility—views and practices of

scholarly publishers. (23 February 2009; www.alpssp.org/ForceDownload.asp?id=129)

Beit-Arie O, Blake M, Caplan P, Flecker D, Ingoldsbey T, Lannom LW, Mischo WH, Pentz E, Rogers S, Van de Sompel H. 2001. Linking to the appropriate copy. *D-Lib Magazine* 7. doi:10.1045/september2001-caplan

Blagoderov V, Brake I, Mayo S, von Raab-Straube E, Rycroft S, Walley L. 2008. IPR and the Web: Challenges for taxonomy. (23 February 2009; <http://editwebrevisions.info/content/meeting-report>)

Cassey P, Blackburn TM. 2006. Reproducibility and repeatability in ecology. *BioScience* 56: 958–959.

Costello MJ, Vanden Berghe E. 2006. Ocean biodiversity informatics: A new era in marine biology research and management. *Marine Ecology Progress Series* 316: 203–214.

Costello MJ, et al. 2008. SCOR/IODE Workshop on Data Publishing. Workshop Report no. 207. (23 February 2009; www.scor-int.org/Publications/wr207.pdf)

Couzin J. 2006. Don't pretty up that picture just yet. *Science* 314: 1866–1868.

Edwards JL, Lane MA, Nielsen ES. 2000. Interoperability of biodiversity databases: Biodiversity information on every desktop. *Science* 289: 2312–2314.

Eysenbach G. 2006. Citation advantage of open access articles. *PLoS Biology* 4: 692–698.

Freeman N, Boston T, Chapman AD. 1998. Integrating internal, intranet and Internet access to spatial datasets via ERIN's Environmental Data Directory. Proceedings of the 26th Annual Conference of AURISA, 23–27 November 1998. AURISA.

Froese R, Lloris D, Opitz S. 2004. The need to make scientific data publicly available—concerns and possible solutions. Pages 268–271 in Palomares MLD, Samb B, Diouf T, Vakily JM, Pauly D, eds. *Fish Biodiversity: Local Studies as Basis for Global Inferences*. Office for Official Publications of the European Communities. ACP-EU Fisheries Research Report 14.

Glover DM, Chandler CL, Doney SC, Buesseler KO, Heimerdinger G, Bishop JKB, Flierl GR. 2006. The US JGOFS data management experience. *Deep-Sea Research II* 53: 793–802.

Heidorn PB. 2008. Shedding light on the dark data in the long tail of science. *Library Trends* 57(2): 280–289. doi:10.1353/lib.0.0036

Howe D, et al. 2008. The future of biocuration. *Nature* 455: 47–50.

Kinne O. 1999. Electronic publishing in science: Changes and risks. *Marine Ecology Progress Series* 180: 1–5.

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

Kohnke D, Costello MJ, Crease J, Folack J, Martinez Guingla R, Michida Y. 2005. Review of the International Oceanographic Data and Information Exchange (IODE). (23 February 2009; www.iode.org/index.php?option=com_oec&task=viewDocumentRecord&docID=336)

Mayo SJ, et al. 2008. Alpha e-taxonomy: Responses from the systematics community to the biodiversity crisis. *Kew Bulletin* 63: 1–16.

McMahon RC. 1996. Cost recovery and statistics Canada. *Government Information in Canada* 2: 4.3. (23 February 2009; www.usask.ca/library/gic/v2n4/mcmahon/mcmahon.html)

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

Minelli A. 2003. The status of taxonomic literature. *Trends in Ecology and Evolution* 18: 75–76.

Orme ER, Jones AC, White RJ. 2008. LSID deployment in the Catalogue of Life. Paper presented at the BNCOD 2008 Biodiversity Informatics Workshop; 10 July 2008, Cardiff University, United Kingdom. (23 February 2009; <http://biodiversity.cs.cf.ac.uk/bncod/OrmeJonesAndWhite.pdf>)

Page RDM. 2006. Taxonomic names, metadata, and the semantic web. *Biodiversity Informatics* 3: 1–15.

Parr CS. 2007. Open sourcing ecological data. *BioScience* 57: 309–310.

[RIN] Research Information Network. 2008. To Share or Not to Share: Publication and Quality Assurance of Research Data Outputs. Research Information Network. (23 February 2009; www.rim.ac.uk/data-publication)

Roberts DM, Rycroft SD, Brake I, Harman K, Scott B, Smith VS. 2008. Getting taxonomy onto the Web. *The Systematist* 30: 3–10.

- Santos C, Blake J, States DJ. 2005. Supplementary data need to be kept in public repositories. *Nature* 438: 738.
- Scholes RJ, Mace GM, Turner W, Geller GN, Jürgens N, Larigauderie A, Muchoney D, Walther BA, Mooney HA. 2008. Toward a global biodiversity observing system. *Science* 321: 1044–1045.
- Smalheiser NR. 2002. Informatics and hypothesis-driven research. *EMBO Reports* 3: 702.
- Testa J. 2004. The Thomson scientific journal selection process. (23 February 2009; http://thomsonreuters.com/business_units/scientific/free/essays/journalselection)
- Weiss P. 2002. Borders in Cyberspace: Conflicting Public Sector Information Policies and Their Economic Impact. US National Weather Service. (23 February 2009; www.weather.gov/sp/Borders_report.pdf)

Mark J. Costello (e-mail: m.costello@auckland.ac.nz) is an ecologist with the Leigh Marine Laboratory of the University of Auckland, Warkworth, New Zealand. He is involved in the use of biodiversity informatics to provide online resources that publish data sets and species information, such as the Global Biodiversity Information Facility, the Ocean Biogeographic Information System, and the European and World Registers of Marine Species.