SCIENTIFIC DATA IN THE PUBLIC DOMAIN

THE NEED TO MAKE SCIENTIFIC DATA PUBLICLY AVAILABLE – CONCERNS AND POSSIBLE SOLUTIONS

LE BESOIN DE RENDRE LES DONNEES SCIENTIFIQUES PUBLIQUEMENT ACCESSIBLE – PREOCCUPATIONS ET SOLUTIONS POSSIBLES

Rainer Froese
IfM-GEOMAR, Düsternbrooker Weg 20, 24105 Kiel, Germany; Tel.: +49431 600 45 79; Fax: +49 431 600 16 99; E-mail: rfroese@ifm-geomar.de

Domingo Lloris
Institut de Ciènces del Mar, P. Maritim de la Barceloneta, 37-39, SP-08003, Barcelona, Spain; E-mail: lloris@icm.csic.es

Silvia Opitz
IfM-GEOMAR, Düsternbrooker Weg 20, 24105 Kiel, Germany; E-mail: sopitz@ifm-geomar.de

ABSTRACT

The paper argues the necessity to render scientific data available in the public domain in order to prevent loss of knowledge associated with institutional discontinuities and poor archiving and conversely to support higher level analyses of biodiversity and ecosystems, often beyond the original scope of data collection. The concerns of data custodians are discussed, e.g. loss of competitiveness, publication by others, copyright and public acceptability of interpretations. Among the solutions suggested to address these are e.g. delayed public access, aggregation of data; proper use agreement and read-only access. It concludes that such public access policy should be in place for all scientific data collected with public funding.

RESUME

Ce travail souligne la nécessité de rendre les données scientifiques publiques afin d'éviter la perte des connaissances suite à des discontinuités institutionnelles et des faiblesses d'archivage, mais aussi afin de permettre des analyses plus poussées sur la biodiversité et les écosystèmes, souvent au-delà de ce qui avait été l'objectif initial de l'échantillonnage. Les préoccupations des gardiens de données sont examinés, telles que la perte de compétitivité, le risque de publication par d'autres, le droit d'auteur et l'acceptabilité publique des interprétations. Parmi les solutions proposées en vue de les prendre en compte figurent le retard d'accès public, l'aggrégation des données, des accords appropriés d'utilisation et un accès limité à la lecture. La conclusion est qu'il serait souhaitable d'avoir des politiques en place sanctionnant de telles solutions pour toutes les données scientifiques collectées avec des fonds publics.

INTRODUCTION

Traditional individual or institutional accumulation of scientific data has caused great losses of knowledge over time, due to lack of long-term archiving and accessibility by both the scientific community and society at large (Pauly, 2001; Zeller *et al*,. 2004). Generally, the fact that data can be used for many more and often unforeseen purposes has largely been ignored. This is especially true for the biological sciences and may explain the lack of effective international cooperation in this field and thus, e.g., the lack of global ecosystem models, in striking contrast to oceanography and climatology, where data sharing and archiving has a long tradition and where results from global models are widely used and accepted (see e.g. Dittert *et al*., 2001; Froese and Reyes, 2003). The Committee on Data for Science and Technology (CODATA) of the International Council of Scientific Unions has suggested almost a decade ago that "scientists supported by public money should make their data available without delay after publication", and that "full and open data access means that not only is there no discrimination in data access, but that the cost is within the reach of scientists in all countries" (CODATA, 1995). The success of FishBase (www.fishbase.org), an online database with over 12 million hits per month and over 300 citations in the scientific literature, demonstrates that a well designed information system that is freely available on the Internet can serve science as well as civil society. The Global Biodiversity Information Facility GBIF (www.gbif.org) and the Ocean Biogeographic Information System OBIS (www.iobis.org) are two recent prominent examples of global data sharing in biological sciences. In contrast, the International Council for the Exploration of the Seas (ICES) requires individual scientists to request access to data through the ICES official channels, including detailed descriptions of what they intend to do with the data. Any cost for data extraction has to be covered by the scientist. A recent offer by FishBase to publicly disseminate ICES data was rejected. Two of the reasons given for this can serve as general examples of concerns of data custodians: "… there is a risk that [open access] will interfere with the willingness of parties to submit data…"; "… data are sensitive and therefore need to be treated correctly and not misinterpreted…" A similar request for access to Russian survey data resulted in an answer that only data for non-commercial species may be made available. Data custodians who gathered at a recent workshop in Barcelona, Spain, expressed a variety of additional concerns. The purpose of this contribution is to understand and list these concerns, to present options that appear suitable for addressing them, and to give examples of the advantages that result from sharing of data.

CONCERNS OF DATA CUSTODIANS

Concerns of data custodians relate to control over data, confidentiality issues, potential misuse of data, lack of trust, and lack of resources. Below we list these concerns together with the typical phrase in which they are expressed.

1. Loss of competitiveness: "If others have access to my data I will lose an advantage e.g. when submitting proposals;"
2. Publication by others: "Someone else will publish my data;"
3. Copyright issues, intellectual property rights (IPR): "I will lose ownership and recognition of my work;"
4. Commercial use of data: "Someone else will make money out of my data;"
5. Public acceptability of conclusions: "The public will dismiss analyses and conclusions when they see the errors and problems in the underlying data (without fully understanding the context);"
6. Manipulation or misreporting of data becoming visible, such as catch statistics derived by adding a fixed percentage every year: "Data are only for internal use;"
7. Problematic data becoming visible: "They will see that we sometimes catch protected species such as turtles. They will see that we sometimes fish outside allowed areas;"

8. Confidentiality of original data providers: “Fishers will stop voluntary data provision if confidentiality can be broken;”
9. Data used for deriving fishing quota are considered sensitive: “We can be sued for misinterpretation of the data;”
10. Misuse of data by others: “They will catch the last existing specimens if we tell them where they are;”
11. Misinterpretation of accuracy of data: “They don´t understand the limitations and assumptions;”
12. Errors in the data becoming visible: “We need time to fix the errors first;”
13. Lack of trust: “They will come and catch our fish if we tell them where they are;”
14. Lack of trust and/or communication between data providers and analysers: “They will make it look as if these were their data;”
15. Lack of benefits: “What do I get out of it?”
16. Limited ability to provide data following standardised concepts and formats: “I need someone to re-structure the data;”
17. Additional work load: “I have no staff to deal with this.”

SOLUTIONS

The concerns expressed above are real-world concerns and thus have to be taken seriously. Not all of the raised issues can be resolved completely. However, a number of solutions have emerged that have proven satisfactory to data custodians. These are presented below.

1. Dissemination delay: Several concerns (2, 4, 6, 7, and 10) relate mostly to recent data. Allowing for a dissemination delay of e.g. 3-5 years can address these concerns. Also, laps of time will typically make ‘inappropriate behaviour’ or violation of rules (concern 7) less relevant politically or legally. Important is that release of data after the respective delay is automatic (programmed in the respective database) and not dependent on administrative procedures. Dissemination delay may not be appropriate for data which are time-sensitive; delaying the release of such data may render them of little use, e.g. for regional management purposes or climate change studies. For such data spatial or temporal aggregation may be a better solution;

2. Aggregation of data: Several concerns (7, 10, and 13) related to the misuse of data, e.g. by poachers or illegal fishers. Allowing for a certain degree of aggregation of sensitive data, e.g. by lumping data in space or time, or blurring exact localities by reporting them in integer degrees. These two approaches are used routinely by, e.g, the U.S. National Marine Fisheries Service for marine fisheries catch data. Such blurring may be permanent if the sensitivity relates to, e.g., occurrence of rare and threatened organisms or spawning aggregations. Data aggregation should be temporary if the sensitivity relates to policy or confidentiality issues, such as what vessel has caught what fish where and when (concern 8); eventually such data should be published with full detail. Again, desegregation after a certain time should be automatic;

3. Data use agreement: Several concerns (1-4, 14, and 15) relate to ownership of data and recognition of work. These concerns can be addressed by making users accept a ‘Data Use Agreement’ which, among other, explicitly states that copyright and intellectual property rights remain with current owners; fair use of data is permitted under the condition that the source of the data is properly cited; and commercial use needs special and explicit permission from the data custodian;

4. Disclaimer: Several concerns (5-7, 11, and 12) related to errors and possible misinterpretation of data. This is a general problem with all data and the standard solution is attaching a disclaimer regarding the quality of data, including full details of the concepts and definitions used,

acknowledgement of known problems, appropriate measures to deal with errors, best tools for proper analysis, and other relevant meta-data;

5. Read-only access: Several concerns (5, 9, and 11) relate to potential manipulation of data by others. Data distributors such as GBIF, OBIS or FishBase have a policy of not modifying data owned by others: adding, editing and deleting is done ONLY by the data provider, who provides new public versions in regular intervals. On the World Wide Web, data are typically extracted on demand from underlying databases; online users are restricted to "read only" access, i.e., while they can download the data and then manipulate them for their own use, they can not temper with the data presented publicly on the web;

6. Respect agreements on confidentiality (concern 8): Confidentiality may be safeguarded by hiding confidential elements of the data-sets such as names of vessels or collectors or by aggregating data as suggested above;

7. Give credit: Several concerns (1-3, 14, and 15) deal with lack of recognition of data providers. FishBase has an internal policy of giving 'more credit than expected' to data providers, such as showing citations, logos and link to web sites of partners in several places. As a result FishBase includes more 'data or photos owned by others' than any comparable information system;

8. Include data owners: Several concerns (1, 2, 14, and 15) relate to the lack of direct benefits for data custodians. Data-dependent projects should include data owners already in the design phase and make sure that data custodians get their due share of support and recognition. However, such projects shall also make explicit when and where data will be made publicly available, and who is responsible for long-term archiving and accessibility. There is an increasing number of projects which have avoided this issue and where data were ultimately lost at great cost to science and society (Zeller *et al*., 2004);

9. Assist data owners: Several concerns (16 and 17) related to lack of capabilities and staff time to make data available. Data distributors such as GBIF, OBIS or FishBase typically provide assistance in form of conversion tools or schemas, or offer to do all necessary conversions themselves.

Advantages of sharing data

As mentioned above, the solutions presented here will not satisfy all aspects of the listed concerns. However, there are a number of advantages resulting from data sharing that typically outweigh any remaining disadvantages and most data custodians who have made their data publicly available will agree that in hindsight that move has been overall beneficiary to themselves and their institutions. For example, FishBase gives very prominent credit to data providers. As a result books that were completely contained in FishBase have sold better than expected and photographers who made their photo collections available have received more requests including from commercial publishers.

Many data owners use public access as a no-cost means of having their data peer-reviewed: feed-back from users helps identifying weaknesses in the data and assists in correcting errors.

Online availability of data usually reduces the number of general requests for information, e.g. in specimen collections, because much of what users need to know is readily available online. On the other hand, visits by specialists—a prime justification for maintaining expensive collections—are increasing as the holdings of collections become better known.

In summary, more exposure typically results in more visibility, recognition, invitations, citations and projects.

CONCLUSIONS

In conclusion we want to stress the principle that scientific data that were established with public funding have to be properly archived and made publicly available. We trust that the concerns of data custodians can be largely addressed by the solutions suggested above. We believe that remaining problems and disadvantages are more than compensated by the advantages resulting from sharing of data.

ACKNOWLEDGEMENTS

We thank David Cross for useful comments on the manuscript and the suggestion to stress advantages of data sharing in a dedicated paragraph. This study was supported by the European Commission, DG Research, within the scope of an INCO Accompanying Measure (ICA4-CT-2002-50001, ECOFISH).

REFERENCES

Dittert, N., M. Diepenbroek and H. Grobe, 2001. Scientific data must be made available to all. *Nature*, 14:393.

Froese, R. and R. Reyes, Jr., 2003. Use them or lose them: The need to make collection databases publicly available. pp. 585-591 *In:* A. Legakis, S. Sfenthourakis, R. Polymeri and M. Thessalou-Legaki (eds.). Proceedings of the 18th International Congress of Zoology.

CODATA, 1995. International Council of Scientific Unions Committee on Data for Science and Technology, CODATA Newletter No. 72 at www.codata.org/newsletters/nl72.html .

Pauly, D., 2001. Importance of the historical dimension in policy and management of natural resource systems, pp. 5-10. *In:* E. Feoli and C.E. Nauen (eds). Proceedings of the INCO-DEV International Workshop on Information Systems for Policy and Technical Support in Fisheries and Aquaculture. *ACP-EU Fish.Res.Rep.*, (8).

Zeller, D. and R. Froese and D. Pauly, 2004. On losing and recovering fisheries and marine science data. *Marine Policy*. [*in press*]