SCOR/IODE Data Publication Project

Data are collected from ocean science activities that range from a single investigator working in a laboratory to large teams of scientists cooperating on large, multinational, global ocean research projects. What these activities have in common is that all result in data, some of which are used as the basis for publications in peer-reviewed journals. However, two major problems remain regarding data: (1) much data that are valuable for understanding ocean physics, chemistry, geology, and biology, and which will help us understand how the ocean operates in the Earth system are never archived or made accessible to other scientists; and (2) when scientists do contribute data to databases, their data become freely available, often with little acknowledgement and no contribution to their career advancement.

The Scientific Committee on Oceanic Research (SCOR) and IODE convened a meeting in Oostende, Belgium on 17-19 June 2008 to discuss how to promote more data submission to approved databases, and better access to these data. A full report of the meeting is available at http://www.scor-int.org/Publications/wr207.pdf. Meeting participants concluded that new infrastructure and new approaches to data publication could help scientists who observe the ocean and model its processes. Most importantly, it is now timely to (1) increase the availability of data used to create figures, tables, and statistical analyses in traditional journal articles; and (2) encourage the expansion of journals that specialize in "data publications" or "data briefs." Data publications are short (as little as a few paragraphs of text) and are designed to describe a data set (not to interpret the data) and provide "persistent" pointers to the data in an approved data repository. Such publications serve an important function of providing a reference that authors can cite on their curricula vitae and should be cited in papers by others who use these data. Journals in the ocean sciences that already welcome such publications include *Marine Micropaleontology*; *Geochemistry, Geophysics, Geosystems; Ecological Archives*, and *Earth System Science Data.* Several other journals also acknowledge the benefits of submitting to approved databases the data underlying traditional publications, as is standard practice in the molecular biology field, in which gene sequences must be submitted to GenBank.

To archive and serve data related to journal publications, an expanded structure in the data management system, data repositories, is required. The purpose of data repositories is to serve as archives of data related to journal articles—both data publications and data backing up traditional journal papers—and serving data that are accessed via persistence identifiers published in the publications. SCOR and IODE will work with existing data centers to promote the development of data repositories at the institutional, national and/or regional level.

A follow-up luncheon meeting was held in December 2008 to bring together a meeting of ocean science journal editors and publishers to discuss how to implement greater use of data publication. The meeting summary follows.

Meeting with Ocean Science Journal Editors
(17 Dec 2008; San Francisco, USA)

Present: David Carlson (*Earth System Science Data* (ESSD)), Dave Checkley (*Fisheries Oceanography*), Peggy Delaney (*Earth and Planetary Science Letters*), Steve Goldstein (Editors' Roundtable), Jim Kirby (*JGR-Oceans*), Roy Lowry (BODC), Ros Rickaby (*Biogeosciences*), Ed Urban (SCOR), Cisco Werner (*Progress in Oceanography*)

Additional written input by Roger Harris (*Journal of Plankton Research*), Ellen Kappel (*Oceanography*), and Eelco Rohling and Jerry Rice (*Paleoceanography*).

Ed Urban explained the purpose of the meeting as to continue the process of consulting with journal editors on the SCOR/IODE data publication activity and to obtain feedback on the SCOR/IODE workshop report. Each participant introduced themselves and gave an explanation of their interest in the idea of data publication and their experience on this topic.

The consensus of the meeting was as follows:

- The effort described in the report is worthwhile and many of the editors consulted want to stay involved in the discussions and participate in the development of the ideas.
- The idea of the peer review of data sets is problematic and will require more discussion and consideration.
- The process of publication of data briefs or other stand-alone papers describing data sets is being tested by ESSD. The SCOR/IODE effort should focus for now on issues related to providing the digital backbone for data related to traditional publications.
- More attention needs to be given to how digital object identifiers (DOIs) can best be used to link journal articles and data sets.
- It would be good to know whether fields outside ocean sciences are pursuing data publishing.

Publication of data associated with traditional journal papers: The scientific communities served by some journals rarely inquire about submitting data or other supplementary information. Earth and Planetary Science Letters (EPSL) requires any data discussed in a paper either to appear in tabulated form in the paper, be readily available in another publication that is referenced in the paper, or be available free from a community-hosted database. EPSL accepts data in Excel tables, which appear as information supplementary to the article, but are only accessible by individuals who subscribe to the article. The data files are stored on the EPSL Web server, but not submitted to any data archives. Usually, data are submitted after the paper is accepted, so that there is a mismatch in timing, in terms of referencing the location of the data in the published paper. Publishing the data behind figures and tables could make meta-analysis easier, without requiring the digitization of data in figures previously published. A system to upload data during the paper submission process (or identifying the data repository in which the data have been placed) would be easy to set up. Perhaps the figure or table legend should be in a standard format from which metadata could be extracted.

Stand-alone data papers: *G-Cubed* publishes data briefs. The first data paper in ESSD will focus on South Pole ozone data and a special issue is planned on ocean carbon data. The ozone paper has triggered publication of related data by other scientists, so the submission of papers may increase quickly. One editor questioned how the creativity of a data set can be assessed to the same degree that the creativity of a traditional journal publication can be judged, implying that maybe data DOIs should not carry the same weight on a person's CV as a traditional paper. Citation of data sets as stand-alone entities could result in data set fragmentation to get a greater number of citations (the" least-publishable unit" problem), but this has not been a problem for ESSD so far and most of the editors who commented supported the idea of stand-alone data publications.

ESSD has three primary criteria for assessing whether they will publish a paper:

1. Fundamental quality of data and metadata
2. Utility of data
3. Archive home identified

ESSD finds data reviewers and follows a process similar to that for a traditional paper. They have identified an initial list of approved data centers, which will be expanded over time.

The Editors Roundtable (a group of editors and publishers of geochemistry journals) has developed a set of common criteria and a checklist for data publications (see Appendix 1).

Data papers might synthesize data that had appeared in previous papers and receive a new DOI because they constitute a value-added product. This results in a data versioning problem. The synthesized data set would not be ingested by a data center because it is already contained in a data center.

One editor had investigated getting DOIs assigned to traditional papers and found it difficult to comply with DOI application requirements.

Reviewing of Data Sets Related to Traditional Publications: It was questioned whether review of data sets associated with traditional journal articles would be feasible, because of the extra effort required. It is hard enough to find reviewers for the normal review process. For some journals, the editor of the journal does a kind of data review, rather than asking the reviewers do it. If data review is established, a checklist should be developed for reviewers. Review of traditional papers already includes a consideration of whether the figures and tables are necessary and scientifically valid. Maybe the data-related review of traditional articles could be as simple as making sure that the data have been deposited in an appropriate data center or repository and that the figure tables and legends could serve as metadata. One editor noted the issue that raw data should be archived somewhere, with metadata containing calibration factors, as it is impossible to evaluate the calibrated data without this information.

“Use cases”: There exist many different use cases (different uses of data publication). It is impossible to solve each use case simultaneously, so it will be necessary to pick the most common uses and develop appropriate procedures for each one, which the SCOR/IODE project will do.

Funding: How will additional data review, publication, and infrastructure be funded?

Data Submission Requirements: One editor suggested that journals should suggest that data associated with articles be submitted to a data center, rather than requiring this, because the culture (or laws) in some countries do not require, support and/or allow data submission. Scientists from such countries would be effectively barred from publishing in a journal that requires data submission.

Appendix 1 – Check list from the Editors Roundtable

CHECK LIST
For the Publication of Geochemical Data

Data

- ❑ **All data listed in data tables**
- ❑ **Downloadable format**
- ❑ **Reference to an accessible, persistent source** such as a public database or data archive *(NOTE: personal web sites are not persistent data archives)*

Comment

Sample Information

• **Unique identifiers:**	❑ Yes	❑ No	❑ N/A
• **Location (lat/long):**	❑ Yes	❑ No	❑ N/A
• **Depth below sea level:**	❑ Yes	❑ No	❑ N/A
• **Depth in core/strat section:**	❑ Yes	❑ No	❑ N/A
• **Classification:**	❑ Yes	❑ No	❑ N/A
• **Age:**	❑ Yes	❑ No	❑ N/A

Comment

Data Table

• **Units:**	❑ Yes	❑ No	❑ N/A
• **Technique:**	❑ Yes	❑ No	❑ N/A
• **Laboratory:**	❑ Yes	❑ No	❑ N/A
• **Reference material(s):**			
Name:	❑ Yes	❑ No	❑ N/A
Measured value(s):	❑ Yes	❑ No	❑ N/A
Number of measurem.:	❑ Yes	❑ No	❑ N/A
Reproducibility:	❑ Yes	❑ No	❑ N/A

Comment

Update

1. 9-11 March 2009: Meeting at IODE to develop use cases for digital publication and determine pilot projects to test them. A few weeks ago, Roy Lowry and Gwen Moncoiffe from BODC, Peter Wiebe from BCO-DMO, Michael Diepenbroek from WDC-MARE, and Peter Pissierssens from IODE met in Oostende to develop two use cases for data publication.

a. Development of citable entities from data ingested into a common schema - retrospectively building citable datasets from 'data atoms'. The idea here is to provide a mechanism to provide rewards for the 'good guy' project participants who hand over their crown jewels to project data management.

b. Providing digital backbone for the conventional paper publication process, storing table data, figure data, syntheses developed for papers and so on, making them publicly available. Could take in data like that associated with EPSL, as well getting rid of the need for journal subscriptions to access these data.

2. Meeting in Woods Hole to develop use case "b" above. The WHOI library has some funding to develop data publication ideas and could take this on. They are planning to meet in March-April.

3. Follow-up with editors and publishers. We will continue to work with the publishers on the issues related to data publication. WDC-MARE is working with Elsevier on this issue.