EOS
EOS, TRANSACTIONS, AMERICAN GEOPHYSICAL UNION

VOLUME 93 NUMBER 43
23 OCTOBER 2012
PAGES 425–436

Pilot Projects for Publishing and Citing Ocean Data

PAGES 425–426

Linking published scientific results with the data on which they are based has been a growing trend. In some disciplines, such as molecular biology, journals require submission of data to a recognized data center as a condition for publication of the associated article. Data centers, government agencies, and journals have been seeking new ways to link publications and data. The push for transparency of science is also moving most fields of science in this direction. The benefits of publishing data and standardizing their provenance have been documented in several locations [*European Union*, 2010; *Ball and Duke*, 2012].

In the ocean sciences, a project was started in 2008 to bring together scientists, data managers, and library experts to explore means to (1) increase the submission of data to data centers, (2) make data more accessible for reuse, (3) link data more closely to traditional journal publications, and (4) create a system that gives more credit to data generators. This project is a joint effort among the Scientific Committee on Oceanic Research, the International Oceanographic Data and Information Exchange (IODE) of the United Nations Educational, Scientific and Cultural Organization's Intergovernmental Oceanographic Commission, and the Marine Biological Laboratory Woods Hole Oceanographic Institution (MBLWHOI) Library (see http://www.iode.org/datapublishing).

The launch of this activity was described in an article in *Eos* [*Lowry et al.*, 2009]. The project is currently conducting two pilot projects to test ideas related to (1) linking traditional journal articles with data and (2) publication of "data snapshots" by data centers.

Linking Traditional Journal Publications With Data

For traditional peer-reviewed scientific articles in the ocean sciences, it is often difficult to find the data on which the article is based, making it difficult to replicate the study, compile the data into a larger analysis, or compare new results with previous results. Most journals encourage or at least allow submission of data in supplemental files, but these may be stored behind firewalls accessible only to journal subscribers and/or available in formats that are not machine readable.

At least one open-access data journal has been created to address this issue, *Earth System Science Data* (ESSD), which has published articles related to ocean sciences. Most ESSD articles provide digital object identifiers (DOIs) that point to data sets held in various data centers, such as PANGAEA and the Carbon Dioxide Information Analysis Center. PANGAEA pioneered the use of DOIs for ocean data sets, but other centers have also begun to use them. AGU provides a location for authors publishing in AGU journals to submit supplementary information, including data (see http://www.agu.org/pubs/authors/manuscript_tools/journals/auxiliary_material/index.shtml). An analysis conducted for this brief report showed that less than 4% of the articles published in *Global Biogeochemical Cycles*; *Geochemistry, Geophysics, Geosystems*; *Geophysical Research Letters*; *Journal of Geophysical Research-Oceans*; and *Paleoceanography* over the period from 1996 to 2011 had data submitted, although the submission rate varied by journal and year.

The MBLWHOI Library is leading a pilot project to identify best practices for linking data sets to traditional journal articles and to give clear credit to data generators. This activity is based on the assumption that data directly related to journal articles should be accessible, discoverable, citable, and freely available through the Internet. Accessibility will require submission of data to a data repository that is stable and permanent. Discoverability will depend on the development of community-accepted metadata standards. Citability will depend on the assignment of a persistent identifier, as well as provenance metadata and attribution.

The MBLWHOI Library is working to define standards and workflows that will support funding agency and publisher mandates. The library identified several published papers to test its e-repository model. An important component of this activity is a partnership of the library with the Biological and Chemical Oceanography Data Management Office (BCO-DMO) at WHOI. Tools and processes were developed to automate the ingestion of metadata from BCO-DMO for deposit with data sets into the Woods Hole Open Access Server (WHOAS) institutional repository and assign a DOI. The assignment of a DOI, enabling accurate data citation, has been identified as an incentive for scientists to make their data accessible.

Ideally, data would be submitted to this process before a paper based on those data is published, so the DOI can be cited in the published paper (as in ESSD), but the library has also been working to link previously published papers with archived data sets (in other words, tracking the provenance of data).

An example of a WHOAS record that includes a paper and the related data can be found at https://darchive.mblwhoilibrary.org/handle/1912/4852 for *Lomas et al.* [2008]. Articles in Elsevier journals that are listed in the ScienceDirect database can now include WHOAS data linkages in their ScienceDirect listings.

Publication of "Data Snapshots" by Data Centers

The British Oceanographic Data Centre (BODC) is leading an effort to publish "data snapshots." Data centers are concerned with preserving and making accessible data in their entirety, free from interpretation and optimized for reuse. Hence, data originators are encouraged to deposit their primary data sets with data centers prior to starting further analyses and interpretation. This process of depositing data to data centers is, and should remain, separate from the process of writing and submitting interpretative papers to scientific journals. This is particularly true in observational sciences, where subsets of data collected in the field can be used and/or combined with data from other observations prior to interpretation and journal paper publication.

A distinction must be drawn between data publication and the normal data center activity of data serving. Data serving includes acquiring data and metadata

from the originating scientists, harmonizing metadata and data file format prior to database ingestion, ensuring that metadata are adequate and accurate and that the data are available in appropriate file formats, and making the most up-to-date version of the data available for interested parties. Data publication, however, requires the assignment of persistent identifiers and the guarantee that the data center will maintain an unchanged version of the data for the foreseeable future. As part of this publication process, data centers will be expected to assess the quality of the metadata and the suitability of the file format.

BODC staff members have created the prototype BODC Published Data Library (PDL), which provides snapshots of specially chosen data sets that are archived using rigorous version management. This enables citation in journal articles through the assignment of a DOI in collaboration with the British Library (https://www.bodc.ac.uk/data/published_data_library/). Data sets selected for inclusion in the PDL are those that are likely to have application beyond a single publication. This makes citation of these data sets by researchers easier: Rather than having to link various data sets (sometimes hundreds or thousands of observations from various sources), they can use the prepacked set and link it to their papers. Data sets will include both those already ingested by BODC as well as those that will be ingested in the future. The PDL catalog can be found at https://www.bodc.ac.uk/data/published_data_library/catalogue/.

Next Steps

Work is continuing on these two activities with the objective of further refining the procedures and eventually mainstreaming these services throughout the research and academic community. To achieve the latter goal, the activities will be widely publicized through participation in relevant conferences. In addition, a "cookbook" is planned to enable other libraries and national data centers to establish similar systems, using best practices. IODE will be implementing a Published Ocean Data system to provide a repository for data from scientists who do not have access to a national oceanographic data center. IODE is also surveying its members to discover national activities on data publication. The group will continue to work with the Data Citation Standards and Practices Task Group of the Committee on Data for Science and Technology, of the International Council for Science (see http://www.codata.org/taskgroups/TGdatacitation/index.html).

References

Ball, A., and M. Duke (2012), How to Cite Datasets and Link to Publications, http://www.dcc.ac.uk/webfm_send/525, Digital Curation Cent., Edinburgh.

European Union (2010), Riding the wave: How Europe can gain from the rising tide of scientific data, final report, Brussels. [Available at http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf.]

Lomas, M. W., F. Lipschultz, D. M. Nelson, J. W. Krause, and N. R. Bates (2008), Biogeochemical responses to late-winter storms in the Sargasso Sea, I—Pulses of primary and new production, *Deep Sea Res., Part II*, *56*, 843–860, doi: 10.1016/j.dsr.2008.09.002.

Lowry, R., E. Urban, and P. Pissierssens (2009), A new approach to data publication in ocean Sciences, *Eos Trans. AGU, 90*(50), 484, doi:10.1029/2009EO500004.

—Ed Urban, Scientific Committee on Oceanic Research, University of Delaware, Newark; E-mail: ed.urban@scor-int.org; Adam Leadbetter and Gwenaelle Moncoiffe, British Oceanographic Data Centre, Liverpool, UK; Peter Pissierssens, Intergovernmental Oceanographic Commission Project Office for the International Oceanographic Data and Information Exchange, Oostende, Belgium; Lisa Raymond, Marine Biological Laboratory Woods Hole Oceanographic Institution Library, Woods Hole, Mass.; and Linda Pikula, Miami Regional Library Atlantic Oceanographic and Meteorological Laboratory, and National Hurricane Center, Miami, Fla.