JCDL136p

SCOR/IODE/MBLWHOI Library Collaboration on Data Publication

Lisa Raymond[1], Linda Pikula[2], Roy Lowry[3], Ed Urban[4], Gwenaëlle Moncoiffé[3], Peter Pissierssens[5], and Cathy Norton[6]

1. Woods Hole Oceanographic Institution (WHOI); 2. NOAA Central Library; 3. British Oceanographic Data Centre (BODC); 4. Scientific Committee on Oceanic Research (SCOR); 5. IOC Project Office of IODE; 6. Marine Biological Laboratory (MBL)

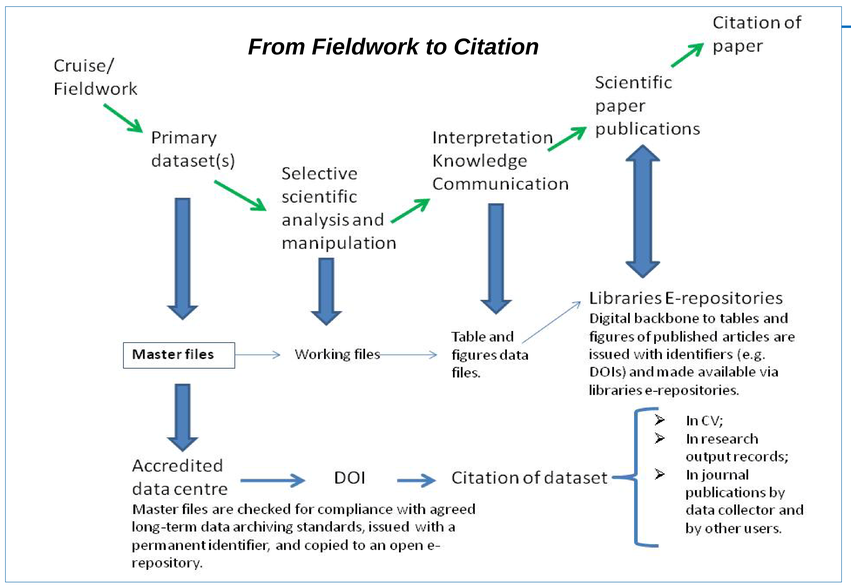


Diagram showing the evolution of a scientific dataset through its different stages from raw data collection (fieldwork) to the generation of master files which must be preserved from accidental loss, self -described, exchangeable, re-usable, and citable (by deposition in an accredited data center), through to the preparation of analyzed data subsets for the purpose of scientific investigation and communication, data subsets which must in turn be openly accessible and citable for the purpose of traceability.

Introduction

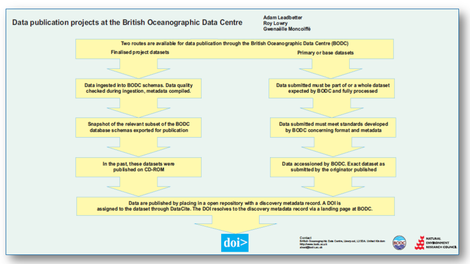
Data collected in ocean sciences, whether generated from research or operational observations, are not always deposited in national or international databases in a format that makes them retrievable and reusable, or even to test the reproducibility of reported research. Often, there are insufficient incentives for data submission, only punishments for not submitting data, resulting in low submission rates and, even when submitted, a bare minimum of metadata. This issue is not unique to the ocean sciences, but several ocean science organizations have begun an effort to stimulate the submission and availability of ocean data.

Method

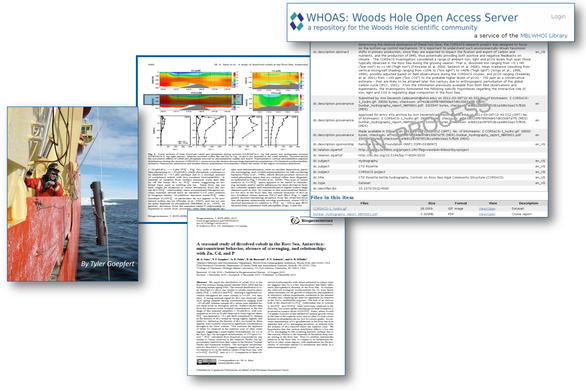
The Scientific Committee on Oceanic Research (SCOR), International Oceanographic Data and Information Exchange (IODE) of the Intergovernmental Oceanographic Commission, and Marine Biological Laboratory/Woods Hole Oceanographic Institution (MBLWHOI) Library are working together to develop and execute pilot projects related to two use cases (1) data held by data centers are packaged and served in formats that can be cited, and (2) data related to traditional journal articles are assigned persistent identifiers referred to in the articles and stored in data repositories, such as DSpace repositories provided by libraries and IODE's PublishedOceanData repository (to be launched in summer 2011). A key feature of this activity is that data centers, libraries, and ocean scientists are working together to test the processes related to these two use cases, and appropriate persistent identifiers will provide a linkage between data, publications, and scientists' CVs.

Use Cases

1. The primary issue under investigation in this use case is how to express the continuous nature of data dynamically managed in a relational database management system in the quantized manner required for citation. The group that is managing this use case (the British Oceanographic Data Centre) is also working to document best practice for the physical composition (e.g. file formats) and semantic description of the content of such snapshots to ensure confident re-use of the data in decades to come.



2. The goal of this use case is to identify best practices for tracking data provenance and clearly attributing credit to data collectors/providers for data published in journal articles. To improve efficacy of data directly associated with a scientific article those data must be discoverable, citable and freely available on the Internet. Resources, standards, and workflows must be defined to support publisher and funding agency mandates. For the data to be discoverable, appropriate metadata, defined using community-accepted metadata standards, must be associated with the data source. Data will be made citable by the assignment of a persistent identifier as well as provenance and attribution metadata. The availability of the data will be assured by submission to a data repository that has stability and permanence.



Conclusions

While authors agree that it would be ideal to begin collaboration early in the research process, we have found that adapting workflows is a challenge. Use case authors have found it difficult to submit data early in the publishing process and QA/QC of datasets is also an issue. Depositing quality data in a timely manner so that DOIs for datasets can be included in the final version of the article continues to be an obstacle. By joining up their effort and expertise, several oceanographic data centers and libraries are putting in place some infrastructure and guidelines which will enable researchers to obtain citable references for their primary datasets prior to publication and also deposit processed data used in journal articles in permanent well managed repositories. Versioning and the management of the relationships between the components of datasets is a key requirement of the data publication process. Recent developments by groups such as DataCite will provide the framework needed to make substantial progress towards the publication of the more dynamic datasets. Collaboration between libraries and data centers such as with the Woods Hole-based data center, Biological and Chemical Oceanography Data Management Office (BCO-DMO), has allowed for deposit of high-quality datasets and procedures are now being developed to automate the system of metadata and data transfer to the repository and the assignment of DOIs. This collaboration is also providing the Library with an opportunity to talk to researchers before papers are published as data issues are being discussed with the data center. In the next year we also expect to renew conversations with publishers as we now have a workable model for oceanographic datasets that can be enhanced by coordination with journal producers.

Related Websites

PublishedOceanData http://publishedoceandata.net/
Woods Hole Open Access Server https://darchive.mblwhoilibrary.org/
Biological and Chemical Oceanography Data Management Office http://www.bco-dmo.org/
British Oceanographic Data Centre www.bodc.ac.uk
Information about the project, including reports from project meetings, can be found at http://www.iode.org/index.php?option=com_content&view=article&id=110&Itemid=129

References

Borgman, Christine L. 2010. *Research Data: Who will share what, with whom, when, and why?* Presented at the China-North American Library Conference, 17 Aug 2010. http://works.bepress.com/cgi/viewcontent.cgi?article=1237&context=borgman

Costello, M. J. 2009. Motivating Online Publication of Data. *BioScience,* 59,5 (May 2009),418-427. DOI= http://dx.doi.org/10.1525/bio.2009.59.5.9

Digital Research Data Curation: Overview of Issues, Current Activities, and Opportunities for the Cornell University Library A report of the CUL Data Working Group. Cornell University, Ithaca, NY, 2008. http://ecommons.library.cornell.edu/bitstream/1813/10903/1/DaWG_WP_final.pdf

SCOR/IODE/MBLWHOI Library Workshop on Data Publication, UNESCO Headquarters, Paris, France, 2 April 2010. IOC Workshop Report No. 230. UNESCO, Paris, 2010. http://www.scor-int.org/Publications/wr230.pdf

SCOR/IODE Workshop on Data Publishing, Oostende, Belgium, 17-19 June 2008. IOC Workshop Report No. 207. UNESCO, Paris, 2008. http://www.scor-int.org/Publications/wr207.pdf

A Woods Hole Data Repository: Addressing the Issues of Provenance, Attribution, Citation, and Accessibility. Project Report. MBLWHOI Library, Woods Hole, MA, 2010. http://lw.rpi.edu/proj/portal.wiki/images/3/3b/JewettSummary.pdf

Acknowledgments

Funding provided by the Jewett Foundation