CORE
Metadata, citation and similar papers at core.ac.uk
Provided by Woods Hole Open Access Server

A WOODS HOLE DATA REPOSITORY: ADDRESSING THE ISUES OF PROVENANCE, ATTRIBUTION, CITATION AND ACCESSIBILITY

Lisa Raymond
MBLWHOI Library
Woods Hole Oceanographic Institution, MS#8
Woods Hole, MA 02543 USA

Abstract: Motivated by publisher and funding agency mandates, as well as a desire to give scientists greater credit for their creation of data, the Marine Biological Laboratory/Woods Hole Oceanographic Institution (MBLWHOI) Library and a team of data managers and scientists are collaborating with the Scientific Committee on Oceanic Research (SCOR) and the International Oceanographic Data and Information Exchange (IODE) of the Intergovernmental Oceanographic Commission 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 institutional repositories, such as DSpace.

Keywords: data management, data publication, metadata, provenance, attribution.

The MBLWHOI Library team has been focused on data that support published articles, particularly the data used to create the figures and tables. The goal was to identify best practices for tracking data provenance and clearly attributing credit to data collectors/providers for data published in journal articles. In order for the data directly associated with a scientific article to be accessible, it needs to be discoverable, citeable and 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 file. Data will be made citeable by the assignment of a persistent identifier as well as provenance metadata and attribution. The availability of the data will be assured by submission to a data repository that has stability and permanence. This paper describes the use of the e-repository model for data publication, the cultural and technical challenges to data deposit and the ongoing collaboration with SCOR and IODE.

The MBLWHOI Library has been working with stakeholders in the management and publication of data with support from the George Frederick Jewett Foundation. In April 2009 a Data Attribution and Provenance for Published Datasets Workshop was held in Woods Hole to gain input from an international group of stakeholders (scientists, data managers and librarians) to determine how to approach a growing problem in science: how to publish data associated with a scientific journal article. The motivation for

publishing data comes from publishers and funding agencies which have new requirements that authors must make available to readers and other interested parties the data underlying the figures, tables and text of submitted manuscripts. The goal was to identify best practices for tracking data provenance and clearly attributing credit to data collectors/providers for data published in journal articles. In order for the data directly associated with a scientific article to be accessible it needs to be (1) discoverable, (2) citable and (3) available on the Internet. Resources, standards, and workflows must be defined to support the publisher and funding agency mandates. For the data to be discoverable, appropriate metadata, defined utilizing community-accepted standards, must be associated with the data file. Data will be made citable by the assignment of a persistent identifier and provenance metadata and attribution. The availability of the data will be assured by submission to a data repository that has stability and permanence.

It was determined during the Data Attribution and Provenance Workshop that the Library would come up with a test case to deposit the data that support figures and tables from an article. The test case would be one of the use cases from the workshop.

The Woods Hole Open Access Server (WHOAS), the MBLWHOI Library's DSpace installation, had already been accepting data, but the new challenge is to integrate metadata fields that were discussed at the workshop and to develop a workflow that will facilitate author submission to journals and incorporate DOI's.

The first task was to review all the metadata tags discussed at the meeting and previous mapping. Workshop participants focused on the Dublin Core and Darwin Core schemas, but also invented some fields, referring to them as Woods Hole Core. The Library staff started with existing fields in WHOAS and then mapped Woods Hole Core to Dublin Core (i.e. wc.process – mapped to dc.description). A couple of Darwin Core fields, dwc.scientificName and dwc.genus were included. More Darwin Core fields can be added, as needed. Two sample dataset records and a record for the draft article were created and the files were loaded on the WHOAS test server for the author to review.

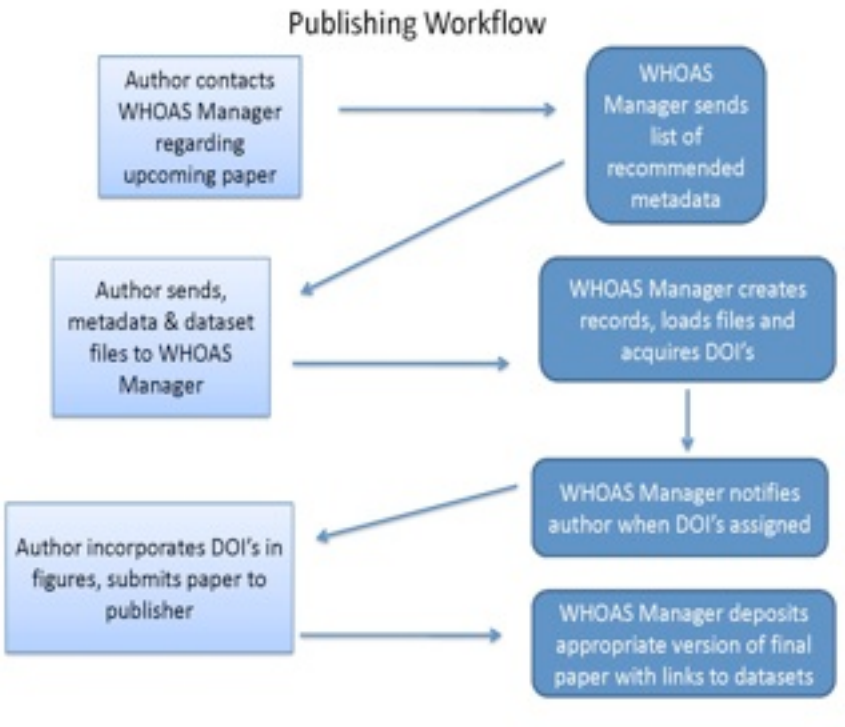

Figure 1. Publishing workflow for processing and storing backbone data associated with a scientific publication.

The next task was to develop a scalable workflow (Figure 1) that incorporates the assignment of DOIs to each dataset for figures and tables. This workflow must be timed to coordinate with the publication process in such a way that DOIs can be included in the final version of a paper submitted for publication.

- Before submitting final draft to publisher, author contacts WHOAS Manager regarding upcoming publication.
- WHOAS Manager informs author of recommended metadata requirements, works with author if other information required.
- Author sends dataset files to WHOAS Manager and provides metadata as necessary.
- WHOAS manager creates records, loads files and acquires DOI's for each dataset.
- WHOAS manager alerts author when DOI's assigned.
- Author incorporates DOI's in figure and tables, submits paper to publisher.

- When the paper is published, the last draft or link to publisher's version will be added to the repository as required by copyright agreement.
- In the article record, links will be added to each dataset record for all figures and tables.

This process ensures that DOIs are assigned in time to be included in the final publication. It could result in assignment of DOIs to datasets for figures or tables that are not included in the final publication. At this time the Library does not see a problem with loading datasets and assigning DOIs to figures and tables that don't get published. The author can decide if the unpublished datasets should have links in the article record. They can remain independent and perhaps even be used for future publication. If individual authors or publishers request a different time line for DOI assignment, we are happy to accommodate their needs. Publishers may eventually incorporate this into their workflow; until then, it was felt that DOIs should be assigned early enough in the process to assure inclusion in publication.

Library staff made a presentation to the author. WHOAS functionality and metadata capabilities were demonstrated and the recommended workflow was presented. The author was pleased with the flexibility of DSpace and the suggested metadata fields, and was happy that the recommended workflow includes data deposit and DOI acquisition early in the publication process. We discussed some issues with DSpace, namely the inability to clearly label links in the article record that point to the figures and tables. Library staff researched this problem and found that the Dryad project has encountered the same obstacle, but continue to add datasets despite this handicap, we have every expectation that future versions of DSpace will address this display issue as Dryad is recognized as a leader in the data deposit arena and they have a strong development team. An advantage of using DSpace is that it generates provenance information in the data ingest process including depositor, date, file size and checksum.

Library staff corresponded with the use case author over the summer and through the fall, requesting the datasets and some additional metadata. However, the paper was accepted for publication and the final draft sent out before the datasets could be submitted and DOI's assigned and included in the publication. The author indicates he would still like to submit the datasets and the Library is ready to accept at any time.

Through the workshop and subsequent work with the use case, several challenges to data publication were identified. Technical challenges, probably the easiest to solve, include defining how much data to include in a backbone dataset, how to deal with multiple proprietary file types and how to preserve deposited data. Cultural challenges include limited incentives for researchers to expend extra resources to publish their data, and fears of theft and loss of control over their data. Cultural challenges will be overcome as funding agencies pressure scientists to make data publicly available. Usual and common challenges to data publication are lack of resources, funding, personnel, and time to publish high quality datasets with adequate metadata. Utilizing and expanding on the information gained from the workshop, additional backbone datasets will be added to the

WHOAS data repository. We have established a workflow and process that will be refined and adapted as we work with scientists from different domains.

During the 2009 IAMSLIC meeting in Brugge Belgium, MBLWHOI Library Director Cathy Norton approached IOC/IODE to have a joint meeting on data publication. In 2008 the Scientific Committee on Oceanic Research (SCOR) and the International Oceanographic Data and Information Exchange (IODE) of the Intergovernmental Oceanographic Commission (IOC) started to work together on the issue of data publication and the development of use cases for data publication began at a meeting in 2009. Use case 1 created data publications from existing and future holdings at national data centers. Use case 2 provided the "digital backbone" for traditional journal publications. On April 2, 2010 the SCOR/IODE/MBLWHOI Library Workshop on Data Publication was held at UNESCO Headquarters in Paris, France. Further developed use cases were presented at the 2010 meeting in Paris and the group agreed that the e-repository model implemented at the MBLWHOI Library was well suited for publication of static data sets in both examples (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 institutional repositories, such as DSpace.

The Meeting recommended that OceanDocs implement a pilot activity following the procedures used in Woods Hole to deposit data. Directions for metadata modifications for Dspace, along with documentation about our workflow, metadata requirements and procedures were sent to IODE shortly after the meeting.

Representatives from SCOR, IODE and the MBLWHOI Library wrote a Data Challenge Document to encourage other libraries and data centers to adopt the e-repository model for data publication. An abstract was submitted to the Committee on Data for Science (CODATA) and a paper will be presented by an IODE representative at the 22nd Annual Meeting in South Africa on the Data Publication Challenge for Ocean Data Management.

A session proposal was submitted to the annual fall meeting of the American Geophysical Union, Earth and Space Science Informatics (ESSI) section wiki on data publication. With collaboration from researchers from Oak Ridge National Lab, Columbia and the Unidata Program Center, a submission was accepted for a session. The session also resulted in an invited speaker session sponsored by the Union. An MBLWHOI Library representative is co-convening the ESSI session and has submitted an abstract for a poster Addressing the Issues of Provenance, Attribution, Citation, and Accessibility.

Further collaboration with SCOR, IODE and Woods Hole based data management programs is expected to be ongoing. The MBLWHOI Library continues outreach to authors as we refine the workflow for data publication. Cultural challenges to data deposit will require innovative approaches, interaction with publishers and support from data centers.

References

SCOR/IODE Workshop on Data Publishing, Oostende, Belgium, 17-19 June 2008. Paris, UNESCO, 23 p. 2008. (IOC Workshop Report No. 207)
http://www.iode.org/index.php?option=com_oe&task=viewDocumentRecord&docID=2457

A Woods Hole Data Repository: Addressing the Issues of Provenance, Attribution, Citation, and Accessibility. Woods Hole, MBLWHOI Library, 15 p. 2010.
http://tw.rpi.edu/proj/portal.wiki/images/3/3b/JewettSummary.pdf

SCOR/IODE/MBLWHOI Library Workshop on Data Publication, UNESCO Headquarters, Paris, France, 2 April 2010. Paris, UNESCO, 11 p. 2010 (IOC Workshop Report No. 230)
http://www.iode.org/index.php?option=com_oe&task=viewDocumentRecord&docID=5437