



United Nations  
Educational, Scientific and  
Cultural Organization



Intergovernmental  
Oceanographic  
Commission

Workshop Report No. 252

# SCOR/IODE/MBLWHOI Library Workshop on Data Publication 5<sup>th</sup> Session

Woods Hole Oceanographic Institution, Woods Hole, USA,  
9-10 October 2012



IOC Workshop Report No. 252  
Oostende, 7 November 2012  
English



Participants of the Workshop (Andrew Maffei participated in the meeting, but is not shown)

**Abstract:**

The fifth SCOR/IODE/MBLWHOI Library Workshop on Data Publication was convened by the Scientific Committee on Oceanic Research (SCOR), the International Oceanographic Data and Information Exchange (IODI) of the Intergovernmental Oceanographic Commission (IOC) and the Marine Biological Laboratory/Woods Hole Oceanographic Institution Library (MBLWHOI Library) on 9-10 October 2012 to evaluate progress of the two pilot projects of the activity and to discuss related topics, such as implementation of data repositories in different data centres and cooperation with related national and international efforts, and hear about how data publication is being handled in other disciplines and interactions with publishers of scientific journals.

For bibliographic purposes this document should be cited as follows:

SCOR/IODE/MBLWHOI Library Workshop on Data Publication, 5<sup>th</sup> Session, Woods Hole Oceanographic Institution, Woods Hole, USA, 9-10 October 2012  
Paris, UNESCO, 7 November 2012 (IOC Workshop Report No. 252) (English)

## **TABLE OF CONTENTS**

|           |                                                                          |                  |
|-----------|--------------------------------------------------------------------------|------------------|
| <b>1.</b> | <b><u>OPENING OF THE MEETING.....</u></b>                                | <b><u>1</u></b>  |
| <b>2.</b> | <b><u>PROGRESS WITH THE USE CASES.....</u></b>                           | <b><u>1</u></b>  |
| 2.1       | USE CASE 1: MBL WHOI LIBRARY/BCO-DMO .....                               | 1                |
| 2.2       | USE CASE 2: BODC-UHASSELT .....                                          | 4                |
| 2.3       | PARTICIPATION IN RCN:OCEANOBSNETWORK .....                               | 5                |
| 2.4       | OUTREACH ACTIVITIES IN 2012 .....                                        | 7                |
| 2.5       | THE DATA CITATION/DATA PUBLICATION COOKBOOK .....                        | 8                |
| <b>3.</b> | <b><u>INVITED SPEAKERS.....</u></b>                                      | <b><u>9</u></b>  |
| 3.1       | CODATA .....                                                             | 9                |
| 3.2       | NISO WORKING GROUP.....                                                  | 11               |
| 3.3       | UPSTREAM PROVENANCE: THE FLEET AS AN OBSERVING SYSTEM .....              | 14               |
| 3.4       | PID INTEROPERABILITY - REPORT FROM IETF AND RDA MEETINGS .....           | 16               |
| <b>4.</b> | <b><u>WORK PLAN 2013.....</u></b>                                        | <b><u>17</u></b> |
| 4.1       | FURTHER DEVELOPMENT OF THE COOKBOOK .....                                | 17               |
| 4.2       | FURTHER TECHNICAL DEVELOPMENTS .....                                     | 19               |
| 4.3       | PROMOTION AND DEPLOYMENT OF THE USE CASES IN THE TARGET COMMUNITIES..... | 20               |
| 4.4       | OUTREACH ACTIVITIES.....                                                 | 21               |
| <b>5.</b> | <b><u>DATE AND PLACE OF NEXT MEETING.....</u></b>                        | <b><u>22</u></b> |
| <b>6.</b> | <b><u>CLOSING OF THE MEETING.....</u></b>                                | <b><u>22</u></b> |

## **ANNEXES**

[ANNEX I: Agenda of the Meeting](#)

[ANNEX II: List of Participants](#)

The Intergovernmental Oceanographic Commission (IOC) of UNESCO celebrated its 50th anniversary in 2010. Since taking the lead in coordinating the International Indian Ocean Expedition in 1960, the IOC has worked to promote marine research, protection of the ocean, and international cooperation. Today the Commission is also developing marine services and capacity building, and is instrumental in monitoring the ocean through the Global Ocean Observing System (GOOS) and developing marine-hazards warning systems in vulnerable regions. Recognized as the UN focal point and mechanism for global cooperation in the study of the ocean, a key climate driver, IOC is a key player in the study of climate change. Through promoting international cooperation, the IOC assists Member States in their decisions towards improved management, sustainable development, and protection of the marine environment.

## 1. OPENING OF THE MEETING

Ms Lisa Raymond welcomed the participants to the Woods Hole Oceanographic Institution and provided information on local arrangements. She then introduced the agenda (attached in Annex I). The list of participants is attached in Annex II.

## 2. PROGRESS WITH THE USE CASES

### 2.1 Use Case 1: MBL WHOI Library/BCO-DMO

This agenda item was introduced by Ms Lisa Raymond and Ms Cyndy Chandler. They recalled the objective of use case 1: Data related to traditional journal articles are assigned persistent identifiers referred to in the articles and stored in institutional repositories. It was noted that the system has been designed such that data sets not associated with journal articles can also be accepted. The use case is based upon the existing Biological and Chemical Oceanography Data Management Office (BCO-DMO) at WHOI that serves as a repository for the storage of data sets. By compiling a metadata summary, making a copy of the data and then creating a record in WHOAS (the library repository) to which a DOI is assigned, the data set becomes a citable reference acceptable to publishers.

Ms Raymond summarized the achievements of Use Case 1 as follows:

- DOIs have successfully been assigned to data associated with articles before and after publication of the article.
- Documentation of procedures in Woods Hole has been completed and sent to Roy Lowry and Adam Leadbetter. The document was made available to the workshop participants by email.
- The system to link from Elsevier's ScienceDirect articles to associated data in WHOAS was implemented.
- The process of metadata and data deposit from BCO-DMO has been automated.
- The ability to assign DOIs, making the data a citeable reference, has prompted conversation with data contributors.
- Very positive responses have been received from the community to presentations and posters presented by BCO-DMO and Library personnel.
- The MBLWHOI Library partnered with the History and Philosophy of Science project (an MBL Center for Library and Informatics grant funded by Arizona State University) to hire a DSpace developer, @Mire, to write code that will allow DSpace to deal with versioning. @Mire has a history of developing code and submitting it to the Dspace community for inclusion in new DSpace releases. The work contracted for, Item Level Versioning, has been submitted for review as part of the next release. The function must be tested by the community, but the MBLWHOI Library is hopeful that it will be accepted and incorporated in DSpace version 3.0, due to be released by the end of the 2012. This capability will allow for a standard procedure to deal with new versions of datasets.

WHOAS (Woods Hole Open Access Server) can be found on <https://darchive.mblwhoilibrary.org>

An issue that has come up is the need to accommodate versioning. A new functionality is being developed in DSpace to allow versioning. It is expected that this will be included in the next version of DSpace. Until that functionality has been added, the MBLWHOI library has adopted an approach that supports versioning of data sets that have been assigned DOIs.

Ms Chandler described the basic dataflow for publishing BCO-DMO data in WHOAS:

1. A dataset is submitted by a scientist to BCO-DMO.

2. BCO-DMO staff enter metadata for the dataset in the BCO-DMO metadata database
3. BCO-DMO staff store the dataset itself in the BCO-DMO data repository.
4. At some later time, a scientist contacts BCO-DMO with a request to cite a dataset.
5. BCO-DMO staff trigger a submission to WHOAS via a special 'submit to WHOAS and request DOI' button in BCO-DMO's administrator's interface to the data system.
6. A utility developed at BCO-DMO pulls data and metadata from the data repository and assembles a METS package.
7. BCO-DMO's database system submits the METS package to WHOAS via SWORD (Simple Web-service Offering Repository Deposit protocol). The DSpace SWORD crosswalk file is an XML file which configures how metadata elements in a mets.xml file are mapped to DSpace metadata elements
8. WHOAS ingests the data automatically and returns a handle URL to BCO-DMO's system.
9. BCO-DMO's system parses the handle URI to determine the corresponding DOI.
10. The DOI is entered in the BCO-DMO metadata database and provided to the scientist who made the citation request.
11. The WHOAS system notifies the WHOAS manager of the submission via email.
12. The WHOAS manager creates a DOI for the submitted dataset record.

Ms Chandler noted that it is also possible to temporarily “embargo” the data, that is, shield them from public view. This is useful when a paper is being prepared. In such a case the DOI is created but the data are not accessible until a set date. So a DOI can be created without the data, simply by entering the metadata.

Ms Chandler concluded her presentation by stating that data associated with published articles can be deposited in institutional repositories and assigned DOIs (even data that still require an embargo period). The assignment of a persistent identifier enables accurate data citation. The ability to cite the data is an incentive for researchers to share their data.

Future plans include submission of a Keyhole Markup Language (KML) file with each data set to provide geospatial context and support a map display of data deposited in the MBLWHOI Library system (BCO-DMO currently sends this as a separate supplemental document).

BCO-DMO can be found on <http://bcodmo.org>

An example of a data object metadata record in WHOAS is shown in [Figure 1](#). An example of the BCO-DMO data object with DOI is shown in [Figure 2](#).

The meeting asked how scalable the system is, if the use of the system by scientists in the Woods Hole community continues to increase. Ms Chandler expected the service to become more popular, but the system is only semi-automated. The submission process is moderated by an information manager at BCO-DMO (the data repository), and on the library side when the DOI is assigned. Moderation by knowledgeable information managers helps to ensure consistent quality of the records. Ms Chandler explained that if the number of submissions increases considerably, the system would need to be modified such that scientists could request DOIs through an online web form.

It was mentioned that California Digital Library offers a service for assigning DOIs, but no details were available.

A question was raised regarding the file size limit. It was noted that at this time no size limit had been set. Data are stored locally. This may need to change in the future. In this regard, it was noted that the URL to which the DOI resolves could always be changed.

Ms Chandler recalled that the initial objective of this group was to have, associated with the paper, all data used for a paper and at a level of granularity that would have provided the “backbone data” for every figure and table in the manuscript, but it was realized that this was not a practical approach. Mr Maffei stated that we might find it advantageous to define what characteristics a “publication DOI”

should have, and perhaps identify other ‘classes’ of DOIs as well, for example, “source data DOIs” and “data product” DOIs and the relationships between them.

Dr Urban asked how the system could be implemented at another location. It was recognized that the planned Cookbook (see below) was a first step. Ms Chandler said that the essential concept was to deposit documented data sets into a permanent archive thereby qualifying them for DOI assignment. It is clear that this process is best done as a joint undertaking between a data centre and library.

The meeting noted further that the Cookbook would need to be a dynamic product that could be further developed based on feedback.

BCO-DMO does some QC on the data but the focus of QC is on the metadata. BCO-DMO counts on the PI to do the QC of the observations and measurements reported in the data set. BCO-DMO sends the data to the US-NODC for permanent archive. No changes are made to the data set when it is archived at US NODC. However, for T-S profiles NODC does do further QC of data sets that are added to World Ocean Database. It was unclear whether there would then be feedback to the BCO-DMO if quality issues were detected in a data submission.

The screenshot displays the WHOAS (Woods Hole Open Access Server) interface. The header includes the WHOAS logo and navigation links. The main content area shows a metadata record for a dataset titled "Biogeochemistry Data collected from the R/V Okeanos cruises : OC399-03, OC408-01 and OC408-02 from the Northwestern Sargasso Sea roughly 35-28°N and 58-68°W, in water depths exceeding 4200 meters, from February 14, 2004 to March 14, 2005 (New Production project)". The record includes fields for Title, Author, Abstract, Citable URI, Date, Created, Location, and Related Material/Data. Below the metadata, there is a section titled "Files in this item" which lists several files with their sizes, formats, and descriptions. The files include a CSV file, a PDF description, a PDF field names, and three KML files for different cruises. The interface also features a sidebar with search and browse options, and a footer indicating the dataset is part of the "Biological and Chemical Oceanography Data Management Office" collection.

**WHOAS: Woods Hole Open Access Server**  
a repository for the Woods Hole scientific community

Search WHOAS  
Advanced Search

Browse  
• All of WHOAS  
• Communities & Collections  
• By Issue Date  
• Authors  
• Titles  
• Subjects  
• This Collection  
• By Issue Date  
• Authors  
• Titles  
• Subjects  
My Account  
• Login  
• Register  
Statistics  
• View Statistics

» WHOAS Home » Woods Hole Oceanographic Institution » Science Data Sets » Biological and Chemical Oceanography Data Management Office » View Item »

Biogeochemistry Data collected from the R/V Okeanos cruises : OC399-03, OC408-01 and OC408-02 from the Northwestern Sargasso Sea roughly 35-28°N and 58-68°W, in water depths exceeding 4200 meters, from February 14, 2004 to March 14, 2005 (New Production project)

Show full item record

**Title:** Biogeochemistry Data collected from the R/V Okeanos cruises : OC399-03, OC408-01 and OC408-02 from the Northwestern Sargasso Sea roughly 35-28°N and 58-68°W, in water depths exceeding 4200 meters, from February 14, 2004 to March 14, 2005 (New Production project)

**Author:** Lomas, Michael W.; Bates, Nicholas R.; Knap, Anthony; Lipschultz, Fredric; Nelson, David M.

**Abstract:** This New Production During Winter Convective Mixing Events (New Production) Project biogeochemistry dataset includes the following data: nutrients, dissolved oxygen, organic matter and alkalinity. Detailed methods for all data collected as part of this study can be found in one of four publications arising from this study. The references include information on analytical machines and certified standards where applicable. The references are listed in the complete dataset description in the supplemental file Dataset\_description.pdf.

**Citable URI:** <http://hdl.handle.net/1912/4853>

**Date:** 2011-10-17

**Created:** 2009-12-07

**Location:** Northwestern Sargasso Sea roughly 35-28°N and 58-68°W  
westlimit: -67.9320; southlimit: 20.6933; eastlimit: -57.1640; northlimit: 35.0337

**Related Material/Data:** <http://hdl.handle.net/1912/4852>

**Files in this item**

| Files                                        | Size    | Format                       | View                       | Description                                                                  |
|----------------------------------------------|---------|------------------------------|----------------------------|------------------------------------------------------------------------------|
| <a href="#">Biogeochemistry.csv</a>          | 434.0Kb | CSV file                     | <a href="#">View/ Open</a> | Biogeochemistry data                                                         |
| <a href="#">Dataset_description.pdf</a>      | 16.21Kb | PDF                          | <a href="#">View/ Open</a> | Additional file information: Dataset description                             |
| <a href="#">Field_names.pdf</a>              | 25.35Kb | PDF                          | <a href="#">View/ Open</a> | Additional file information: Field names                                     |
| <a href="#">OC399-03_Biogeochemistry.kmz</a> | 95.95Kb | application/vnd.google-earth | <a href="#">View/ Open</a> | Biogeochemistry data (zip-compressed KML files to be viewed in Google Earth) |
| <a href="#">OC408-01_Biogeochemistry.kmz</a> | 34.70Kb | application/vnd.google-earth | <a href="#">View/ Open</a> | Biogeochemistry data (zip-compressed KML files to be viewed in Google Earth) |
| <a href="#">OC408-02_Biogeochemistry.kmz</a> | 35.99Kb | application/vnd.google-earth | <a href="#">View/ Open</a> | Biogeochemistry data (zip-compressed KML files to be viewed in Google Earth) |

This item appears in the following Collection(s)  
• Biological and Chemical Oceanography Data Management Office  
BCO-DMO

Figure 1: Data object metadata record at WHOAS

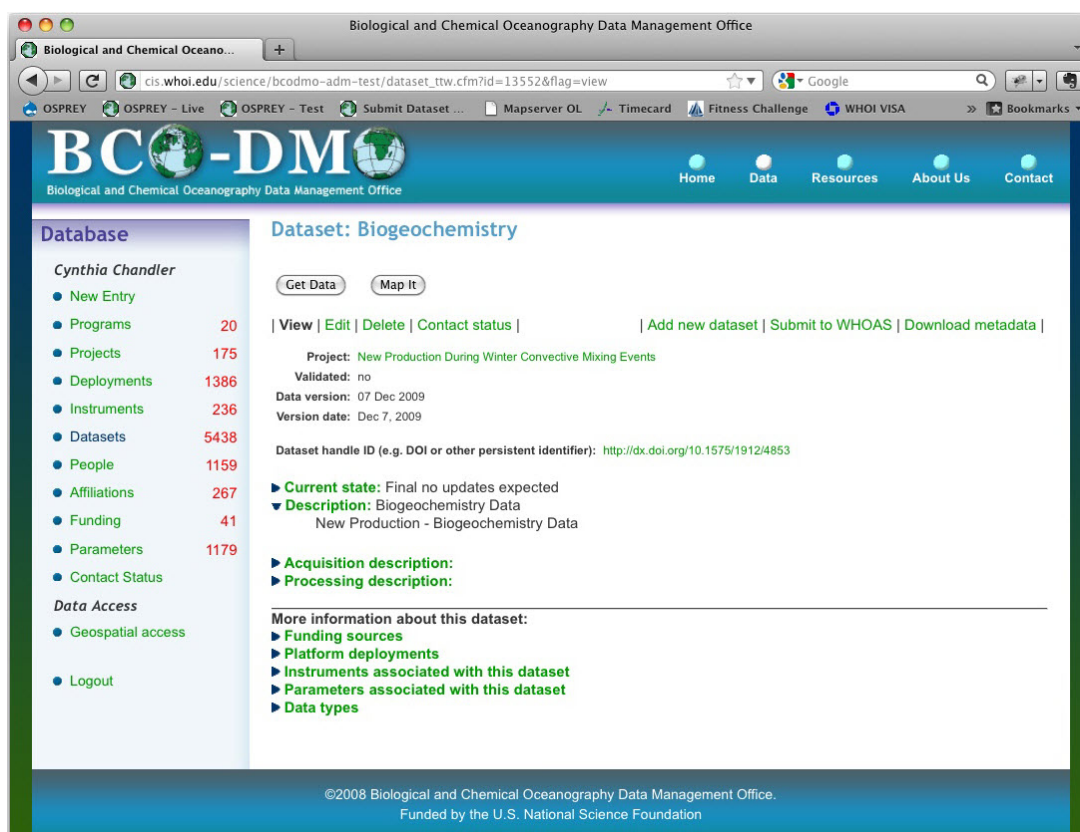


Figure 2: BCO-DMO Data Object with DOI

## 2.2 Use Case 2: BODC-UHasselt

This agenda item was introduced by Dr Adam Leadbetter. He noted that the presentation had been prepared jointly with Dr Roy Lowry and Dr Gwen Moncoiffé, who unfortunately were unable to attend the meeting.

He started his presentation by recalling the objectives of Use Case 2 and in particular the BODC's Published Data Library (PDL): the delivery of meaningful collections of data. Data delivered by the PDL must be (i) fixed to the checksum level; (ii) discoverable; (iii) usable with confidence without referral to any additional material (Context!); and (iv) assured to be available for the foreseeable future.

So the objective here is to work with certain data sets that were, for example, outcomes of projects and that are made available as a “product”.

Dr Leadbetter stressed that the PDL is not a replacement for BODC's “traditional” data serving. It is a parallel system tailored to suite the needs of the academic publishing community. It does not support “data behind the graph”. So it is a data set for a paper, not for the individual elements (tables, graphs) in the paper, as was originally intended, but proved unrealistic. The meeting noted that BCO-DMO had come to the same conclusion.

Within PDL, data sets are assigned DOIs through DataCite (see <http://www.datacite.org>) (NERC, the parent body of BODC, is affiliated to DataCite through the British Library). The DOI prefix is 10.5285 and the suffix is a UUID. The DOIs resolve to an html landing page which contains metadata related to the data set. The landing page links to usage metadata and to the data.

To each metadata record multiple data files are attached. They are not zipped together. Also ODV files are used to allow combining, for example, several profiles. The files are linked to static versions of the data, not a version in the BODC data system. So this may be a static copy of a CD.

At the date of the workshop the descriptive pages of PDL are live on  
[https://www.bodc.ac.uk/data/published\\_data\\_library](https://www.bodc.ac.uk/data/published_data_library)

In addition the DOI catalogue is live (including 8 data sets with DOIs):  
[https://www.bodc.ac.uk/data/published\\_data\\_library/catalogue](https://www.bodc.ac.uk/data/published_data_library/catalogue)

For each of the 8 data sets with DOIs, the landing page is available, showing the metadata in the repository as well as links to the data set and the usage metadata. The landing pages contain human- and machine-readable metadata. They use html and RDFa. Updates of the data are noted as hAtom (a microformat for content that can be syndicated) entries. The fields are documented in the Cookbook (a copy of the draft cookbook was made available to the workshop participants).

#### Future work:

Dr Leadbetter noted that the PDL currently is entirely hand-coded. A design has been documented for an RDBMS back office and preliminary population has commenced. Links with the IODE POD (Published Ocean Data) will be established before the December 2012 AGU Fall Meeting.

The PDL system has been announced to the UK science community through the 5 environmental data centres in the UK (British Oceanographic Data Centre, Centre for Environmental Data Archive, Polar Data Centre, Environmental Information Data Centre).

Ed Urban suggested that a presentation about the PDL at the 2014 Challenger Society meeting could help make it more visible within the UK oceanographic community.

Dr Leadbetter also introduced both the *Geoscience Data Journal* and the JISC-funded PREPARDE project, in order to inform the group of other activities on data publication within the United Kingdom's environmental science community.

The *Geoscience Data Journal* (GDJ) is supported by NERC to publish data and data papers. It is open access. It was launched in April 2012 and is accepting submissions now.

Differences between the two systems (WHOAS BCO-DMO and PDL): we started out from slightly different positions, but technically the systems are very similar. The metadata structures (Dublin core-based) are the same.

So we need to abstract the services required so these can be developed within existing systems. Dr Leadbetter informed the meeting that discussion had been held considering implementing DSpace and EPrints for PDL but these were rejected. It was then decided to custom build the required application.

## **2.3 Participation in RCN:OceanObsNetwork**

This agenda item was introduced by Ms Lisa Raymond and Mr Peter Pissierssens. They explained that the Research Coordination Network (RCN): OceanObsNetwork is a project funded by the U.S. National Science Foundation. The goal of the RCN is to foster a broad, multi-disciplinary dialogue, enabling more effective use of ocean observing systems, consistent with national and international efforts, to inform societal decisions. In achieving these objectives, the RCN will motivate new research outcomes, provide wider visibility for the value and impacts of ocean observations and encourage a new generation of scientists to focus on the oceans and their challenges. The RCN is primarily a forum to address issues of enhancing ocean observation and information. It is not a body chartered to undertake new scientific research. Issues engaged by the RCN will be addressed by the body as a whole (Plenary) or through working groups (WG) constituted by the RCN. A working group

will generally focus on one of the objectives cited above and will produce a report clearly identifying the issues, approaches, impacts and recommendations for achieving the objective(s).

The OceanObsNetwork goals and objectives are to foster a broad, multi-disciplinary dialogue, enabling more effective use of sustained ocean observatories and observing systems. To achieve these, the activities for the RCN in the following areas are planned with the specific actions addressed in the subsection below:

1. Approaches for Sustainability of Observing Systems
2. Stimulating interdisciplinary cooperation
3. Facilitating Open Exchange of Data and Information
4. Promote Standards and Interoperability including approaches for common sensor builds
5. Improving the flow of critical information to key stakeholder
6. Capacity Building
7. Outreach and Community Building

IODE was invited to participate in the RCN:OceanObsNetwork to contribute to activity 3 “Facilitating Open Exchange of Data and Information”. Under this activity it was stated *“While there are many technical issues for open access (see next sections), the policy and cultural issues, even within the ocean/academic community will dominate discussions. For example, SCOR and IODE are looking at the challenges of career advancement for publishing quality data without interpretive analyses. Historically, data is sequestered, sometime for years, while preparing analyses and publication. Such items will be points of dialogue for the RCN in conjunction with other organizations. Other areas of intellectual property and national security are less tractable for the science community. Policy aspects of the free and open access issue has been take up by GEO at the ministerial level. The RCN contribution will be to address issues within the context of globalizing ocean observations and input to discussions that take place through GEO, IOC, WMO, etc. The RCN will thus encourage the formation of a team to address these areas with a focus on observatories and coastal observations.”*

Observations collected need to be managed and made available to the research community. A balance needs to be found between the interests of the individual or group who was responsible for the collection of the data (and who wishes to use these observations for intellectual work that will contribute to science as well as to his/her career) and those of the global ocean observation and science community.

The working group dealing with "Facilitating Open Exchange of Data and Information" was instructed to address the above mentioned balance by considering the following elements:

1. Data and Information formats and standards
2. Data access models (including intellectual property rights, business models for open data, data policies,...)
3. Data publishing, data citation

Each of these would be dealt with by a Task Team. IODE was asked to deal specifically with item 3: Data publishing and data citation. The IODE Secretariat then contacted members of the SCOR/IODE/MBLWHOI Library project on Data Publication as well as other relevant IODE experts. Lisa Raymond, Pauline Simpson, Amy Stout, Linda Pikula, Marc Goovaerts (by email only) and Peter Pissierssens were then identified as members of the Task Team.

Several Webex meetings were held between May and October 2012. As an outcome of its 3 August 2012 meeting a document was prepared by the Task Team entitled [“Report of the RCN:OceanObsNetwork Task Team on data publication/citation \(8 August 2012\)”](#). This document was reviewed by Albert Williams (WHOI) and Jay Pearlman on 30 September 2012 and is currently being revised by Lisa Raymond and other members of the Task Team. It will be part of a larger

document that will be produced in October. In addition, a meeting of the Working Group is planned to be held during the AGU Fall Meeting in San Francisco, December 2012. This will provide further visibility for the SCOR/IODE/MBLWHOI Library project.

The RCN web site is <http://rcn.iode.org>

The Task Team concluded that data publication that enables data citation can certainly be an incentive to make data accessible. The associated functionality to deposit data safely and securely should be attractive to the researcher and the additional citation of the data associated with a research paper will add value to these data as an essential component of research output.

## 2.4 Outreach activities in 2012

This agenda item was introduced by Ms Linda Pikula. She explained that the newest outreach activity for the data publication project was the OceanTeacher Academy Course Data Management for Marine Information Managers, taught on September 26-30, 2011 in Oostende, Belgium at the IOC Project Office for IODE. The course included participants from 13 countries.

The course had been developed taking into account that there are several training courses devoted to data management and access. However, most seem to be geared toward scientists. As libraries become partners in the management and archiving of research data, existing skill sets must be expanded and new expertise developed. This course aimed to introduce information professionals to the history of scientific data and the scholarly information e-science lifecycle. Practical experience in digitization of a document with inclusion of data link was covered.

Topics covered included (i) Introduction to the information professionals' role in data curation and management; (ii) Survey and assessment of data, data profiles, data management plans and data citation and metadata; (iii) New and developing methodologies, as well as data mandates by scientific funding agencies and governments; (iv) Current Data Projects; (v) Skills, role and career structure of data scientists and librarian curators now and in the future; and (vi) Practical Digitization exercises.

Lecturers included (i) Anthony D. Smith, Senior Program Officer, Institute of Museum and Library Services, U.S. and former Univ. of M\_\_\_? Digital Projects Manager; (ii) Lisa Raymond, Associate Library Director, MBLWHOI Library, Manager WHOI Data Library & Archives; (iii) Linda Pikula, Chief Marine Information Editor, IODE OceanTeacher and NOAA Central Library, Miami Regional Librarian; (iv) Paul Nieuwenhuysen, Professor, Vrije Universiteit Brussel; and (v) Marc Goovaerts, U of Hasselt, Belgium, Library Director.

The course outline is available at <http://classroom.oceanteacher.org/course/view.php?id=146>

Ms Pikula then briefly recalled other outreach activities:

### Posters and Publications

- "Ocean Data Publication and Citation" by listed authors Ed Urban, Adam Leadbetter, Gwenaëlle Moncoiffe, Peter Pissierssens, Lisa Raymond, and Linda Pikula (to be published Oct. 2012) *Eos*
- "Citing with Confidence: The Role of Persistent Identification of Agents and Objects in Research " Authors: Lisa Raymond<sup>1</sup>, Cynthia L. Chandler<sup>2</sup>, Roy Lowry<sup>3</sup>, Ed Urban<sup>4</sup>, Gwenaëlle Moncoiffe<sup>3</sup>, Adam Leadbetter<sup>3</sup>, Peter Pissierssens<sup>5</sup>, Cathy Norton<sup>6</sup>, and Linda Pikula<sup>7</sup> AGU Poster Presentation, Fall, 2012.
- Raymond, Lisa, Linda Pikula, Roy Lowry, Ed Urban, Gwenaëlle Moncoiffe, Peter Pissierssens and Cathy Norton. 2011. *SCOR/IODE/MBLWHOI Library Collaboration on Data Publication*. Peer reviewed poster given at ACM/IEEE, Joint Conference on Digital Libraries June 13-17, 2011, University of Ottawa, Ottawa, Canada.

- Raymond, Lisa, Linda Pikula, Roy Lowry, Ed Urban, Gwenaëlle Moncoiffe, Peter Pissierssens and Cathy Norton. 2011. *SCOR/IODE/MBLWHOI Library Collaboration on Data Publication*. 1st International Council for Science (ICSU) World Data System Conference, Global Data for Global Science, September 3-6, 2011, Kyoto University, Kyoto, Japan.
- Lisa Raymond: Data publication presentations at IAMSLIC, Argentina 2010, Alaska, 2012
- Lisa Raymond: AGU poster 2012
- A. Leadbetter, S. Callaghan, R. Lowry, G. Moncoiffé, S. Donnegan, S. Pepler, N. Cunningham, P. Kirsch, L. Ault, P. Bell, R. Bowie, K. Harrison, B. Smith-Haddon, A. Wetherby, D. Wright, and M. Thorley, 2012. *Data publication activities in the Natural Environment Research Council*. EGU General Assembly 2012 - available through <http://meetingorganizer.copernicus.org/EGU2012/EGU2012-2935.pdf>
- Adam Leadbetter, Roy Lowry, and Gwenaëlle Moncoiffé, 2011. *Data publication projects at the British Oceanographic Data Centre*. EGU General Assembly 2011 – available through <http://meetingorganizer.copernicus.org/EGU2011/EGU2011-2142.pdf>
- L. Raymond, C Chandler, R Lowry, E Urban, G Moncoiffe, P Pissierssens, C Norton, and H Miller. 2012. "Data Publication: A Partnership between Scientists, Data Managers and Librarians". Geophysical Research Abstracts, Vol. 14, EGU2012-1440, poster presented in Session "GI 1.3 Data publishing: paving the way to data-intensive science" at EGU General Assembly 2012, Vienna, Austria. Available through : <http://meetingorganizer.copernicus.org/EGU2012/EGU2012-1440.pdf>
- Raymond, L, C. Chandler, R. Lowry, E. Urban, G. Moncoiffe, P. Pissierssens, C. Norton, H. Miller. 2012. "Emerging Role of Librarians in Data Publication". Poster presented at University of Massachusetts and New England Area Librarian e-Science Symposium. April, 2012, UMass Medical Center, Shrewsbury, MA, USA.  
[http://escholarship.umassmed.edu/escience\\_symposium/2012/](http://escholarship.umassmed.edu/escience_symposium/2012/) ;  
poster PDF:  
[http://escholarship.umassmed.edu/cgi/preview.cgi?article=1030&context=escience\\_symposium&cc=c64m4rj3&login=1654102](http://escholarship.umassmed.edu/cgi/preview.cgi?article=1030&context=escience_symposium&cc=c64m4rj3&login=1654102)

#### Other activities

- Collaboration with RCN: OceanObsNetwork Task Team on Data Publication – August, 2012 ongoing (see also agenda item 2.2)
- Peter Pissierssens and Linda Pikula: IAMSLIC Alaska, August 2012 Data Panel: Data and the Librarian's Role.

The meeting then considered future outreach opportunities. These are reported under agenda item 4.4.

## **2.5 The Data citation/data publication Cookbook**

This agenda item was introduced by Dr Adam Leadbetter. He recalled that the “cookbook” has been written for data managers and librarians who are interested in assigning a permanent identifier to a dataset for the purposes of publishing that dataset online and for the citation of that dataset within the scientific literature. The cookbook outlines what is meant by the term “data publication”, provides a step-by-step guide to the data publication process and showcases some data publication best practices. A draft was made available to the workshop participants before the meeting by email.

The Cookbook explains data publication (what is data publication, why data publication) and describes the data publication process (assigning a persistent identifier, metadata, data). It also provides a few “best practice” examples.

The first draft was circulated to the group on 24 September 2012. Additional content was received from Ms Lisa Raymond. A second draft will be circulated by 26 October 2012, in order to incorporate discussion points from this meeting. The aim is to complete the Cookbook for publication by November 2012.

The meeting then considered whether there are other use cases and addressed the following questions:

- Q: Is the system now too complex?  
A: The simplest structure would be a standard DSpace installation with additional schemas and fields + data storage (which could be as simple as a FTP site).
- Q: Who should be able to use it?  
A: The meeting strongly recommended that the system should be managed jointly by data managers and marine librarians, or one of these if they are not both available in a given institution.
- Q: Who is the audience?  
A: Research groups (projects), research institutions, academic institutions, data centres, libraries associated with the former. But also with outreach to individual scientists.

The meeting briefly considered the need for additional content in the Cookbook and decided to discuss the revised table of contents under agenda item 4.1.

The meeting recommended that the cookbook should be available online but also as a printable PDF. Accordingly it was agreed to develop the Cookbook as a Word (via Google Docs) document and to develop OceanTeacher content for the courses.

It was recommended to create a “data publication page” in Wikipedia as well.

## 3. INVITED SPEAKERS

### 3.1 CODATA

This agenda item was introduced by Mr Helge Sagen. He explained that this presentation had been jointly prepared with input from Paul F. Uhler, Daniel Cohen, Franciel Linares, and Todd Carpenter.

Mr Sagen informed the workshop that he would report on the outcome of the CODATA Task Group on Data Citation Standards and Practices, held at the DataCite meeting in Copenhagen in June 2012. He would also provide some information on recent development within Norway.

Full information on the CODATA Task Group is available on <http://www.codata.org/taskgroups/TGdatacitation/index.html>

Mr Sagen explained that the Task Group is led by 3 Co-Chairs: (i) Jan Brase, (Director, DataCite, and ICSTI representative), Technische Informations Bibliothek (TIB)/German National Library of Science and Technology, GERMANY; (ii) Sarah Callaghan (U.K. CODATA), The NCAS British Atmospheric Data Centre, STFC Rutherford Appleton Laboratory, UNITED KINGDOM; and (iii) Bonnie Carroll (U.S. CODATA and CENDI), President, Information International Associates, USA.

The objectives and expected deliverables are as follows:

- Conduct inventory and analysis of existing literature and existing data citation and attribution initiatives.
- Investigate and analyse how existing data repositories cite and provide attribution to their data sets.
- Identify and obtain input from stakeholders in the library, academic, publishing and research communities.
- Provide an international forum to identify and help reconcile the needs of various stakeholder communities.
- Share information and create greater awareness of these issues internationally.
- Establish a public web presence.
- Conduct meetings and workshops to articulate the state of the art and best practices in this area, and to identify emerging issues.
- Work with the major international, regional, and national standards organizations to develop formal data citation and attribution standards and best practices.
- Promote scientific data attribution by developing models, tools, and practical guidance on how to publish citable and traceable data sets.

A symposium and workshop was held in Berkeley, California in August 2011. More information is available on [http://sites.nationalacademies.org/PGA/brdi/PGA\\_063656](http://sites.nationalacademies.org/PGA/brdi/PGA_063656) . The report is expected to be published in October 2012.

The bibliographic inventory is ongoing. Interviews have been organized with a sample of identified stakeholders concerning data citation and attribution practices. These include 28 data repositories, 14 publishers, 9 researchers and 10 funding organizations. It is available through [http://www.codata.org/taskgroups/TGdatacitation/Bibliography\\_Links.html](http://www.codata.org/taskgroups/TGdatacitation/Bibliography_Links.html) page or direct link at [http://www.codata.org/taskgroups/TGdatacitation/docs/CODATA\\_DDCTG\\_BestPracticesBib\\_FINAL\\_17June2012.pdf](http://www.codata.org/taskgroups/TGdatacitation/docs/CODATA_DDCTG_BestPracticesBib_FINAL_17June2012.pdf)

A white paper on current practices in data citation is being prepared (for February 2013)– an outline has been developed. A Best Practice White Paper is planned for 2014.

The next session of the Task Group will meet in Taipei, 27-28 October 2012, during the International Conference “Open Data and Information for a changing planet”, 28-31 October 2012.

During the Conference there will be a sessions related to Data Publication and Data Citation, subdivided in three parts:

- D2 Data Publication and Data Citation (Part 1) - [Co-organized by WDS and CODATA Citation Task Group] - Research Data Enters Scholarly Communication  
*Chair: Mustapha Mokrane*
- E2 Data Publication and Data Citation (Part 2) - [Co-organized by WDS and CODATA Citation Task Group] - Linking Data and Documents: Deepening the Scholarly Record  
*Co-Chairs: Bonnie Carroll & Paul Uhler*
- G2 Data Publication and Data Citation (Part 3) - [Co-organized by WDS and CODATA Citation Task Group] - Roundtable on Data Citation and Practices  
*Co-Chairs: Paul Uhler & Bonnie Carroll*

More on <http://www.codata.org/Taipei/official-program.html>

The DataCite summer meeting was held in Copenhagen, Denmark at the GBIF Offices. The presentations were video-recorded and can be seen on [http://www.youtube.com/playlist?list=PLB8905A7D82CE27EB&feature=view\\_all](http://www.youtube.com/playlist?list=PLB8905A7D82CE27EB&feature=view_all)

Mr Sagen also recalled that in May 2012 the **ISO standard 26324** was published. That means that the DOI is now an ISO standard. ISO 26324:2012 specifies the syntax, description and resolution functional components of the digital object identifier system. It also gives the general principles for the creation, registration and administration of DOI names. See also <http://www.iso.org/iso/news.htm?refid=Ref1561>

Regarding recent developments within Norway, Mr Sagen mentioned “Increased awareness of data citation META 03/2012”. META is the magazine published by the Notur II project, the Norwegian metacenter for computational science. The magazine covers a variety of topics related to e-Infrastructure for computational science in Norway. The magazine contains popular scientific articles, news items, opinions, and information about relevant national as well as international developments. See <http://www.notur.no/publications/magazine/>.

He informed the meeting that the Ministry of Fisheries and Coastal Affairs is now instructing the institute to publish data sets.

The meeting welcomed the development of the bibliographic inventory. □

The meeting regretted that none of the group members had been invited to the Taipei meeting as the topic is of great relevance to data publication.

The meeting requested Mr Sagen to share the Cookbook with the CODATA team.

The meeting requested Mr Sagen to check with CODATA if they would share the white paper with our group. Mr Sagen contacted Mr Paul Uhler during the meeting, who responded that the document was not yet ready so sharing it would be somewhat premature. However, a review process will start in December 2012/January 2013 and our group could possibly participate in that review.

Mr Sagen also requested more frequent interaction with the CODATA Task Group. Mr Uhler agreed that there should be more interaction with other interested groups and some more structured communications/coordination with the SCOR/IODE/MBLWHOI group (and others) would be desirable. Mr Uhler will copy the Task Group co-chairs about this suggestion and proposed to have it as a discussion item in the next Task Group teleconference and at the Task Group meeting in Taipei.

The meeting stressed the importance of reinforcing each other’s activity and invited CODATA to actively share information on its activities. Some of this could be included in the cookbook.

## 3.2 NISO working group

This agenda item was introduced by Ms Cathy Norton. The presentation was given with slides from Amy Stout who served on the committee and Alexander "Sasha" Schwartzman, co-chair of the task force for NISC/NFAIS, Ms Norton presented some slides from the presentation that Schwartzman gave at the CrossRef workshop meeting in Cambridge, Massachusetts on November 14, 2011. Ms Norton explained that the presentation would provide a summary of NISO/NFAIS Recommended Practices for Online Supplemental Journal Article Materials. It was noted that the recommendations in this document are intended to help publishers and editors guide authors and peer reviewers in their work and to provide common ground for delivery of supplemental materials.

The final draft recommendations are available at: <http://www.niso.org/workrooms/supplemental>. The general principles of the document are as follows:

- The recommendations contained in the document are meant to be a guide rather than edicts.
- The document focuses on what future researchers will need rather than previously established paradigms.
- Supplemental materials associated with a publication must be relevant to the primary scholarly work.

- It is up to individual journals to decide what constitutes supplemental materials.
- Descriptive metadata should be associated with supplemental materials in order to clearly explain the contents.
- The recommendations are a living document and should be flexible and change over time.
- The recommendations are mostly for publisher-hosted content, although the guidelines may be applied to repository-hosted content as well.

Supplemental materials should be divided into two general categories: Integral Content and Additional Content.

- Integral Content: Editors should review any materials deemed essential to the full understanding of the article at the same level as the article.
- Additional Content: These materials should be reviewed at the same level as the article, but resources may not be available to do so.

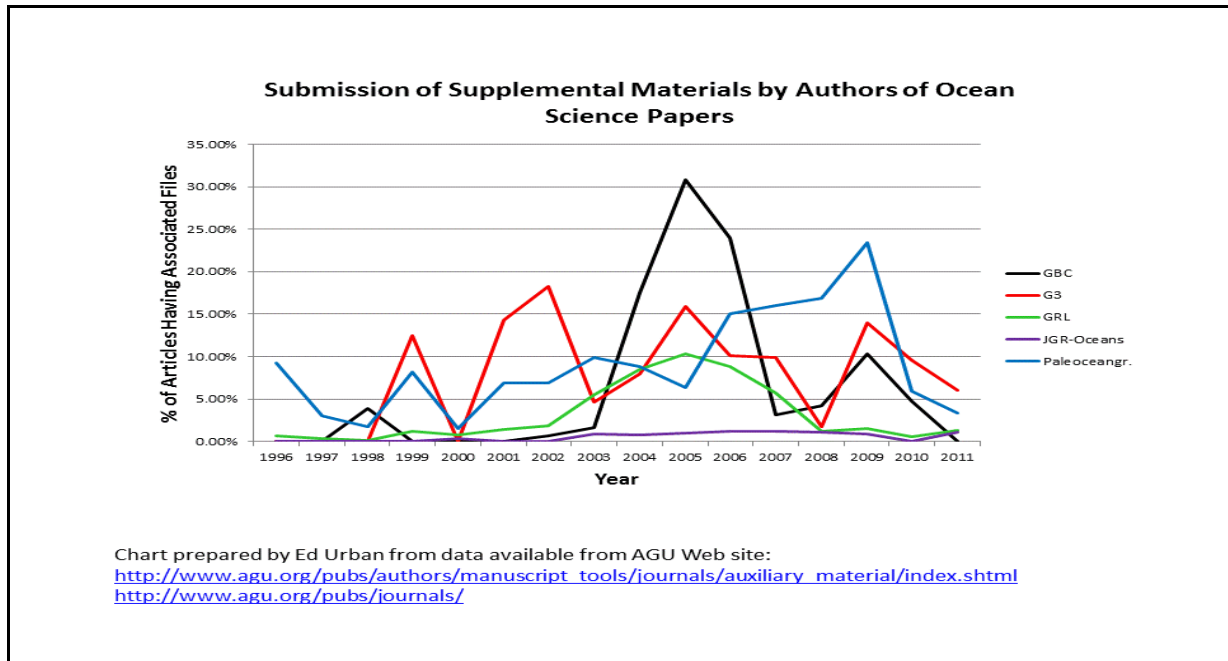
Supplemental materials should be hosted by the publisher, the publisher's host/aggregator, or an external repository. Publishers should use persistent identifiers, preferably DOIs, to identify supplemental materials.

Recommend required metadata elements include:

- Article metadata
- Article DOI
- Supplemental material description
- Supplemental material DOI
- Supplemental material object metadata
- Relationship of supplemental material to article
- Supplemental material physical metadata
- Supplemental material formats

There are also optional metadata elements that may provide a fuller description.

Ed Urban then showed graphs ([Figure 3](#) and [Figure 4](#)) indicating that data publication in ocean journals has not been increasing but has been quite erratic over time. This figure was prepared from data available from AGU Web site: [http://www.agu.org/pubs/authors/manuscript\\_tools/journals/auxiliary\\_material/index.shtml](http://www.agu.org/pubs/authors/manuscript_tools/journals/auxiliary_material/index.shtml) and <http://www.agu.org/pubs/journals/>. He also showed a pie chart showing the file types associated with the journal *Global Biogeochemical Cycles*, indicating that not all of the supplemental files are data files.

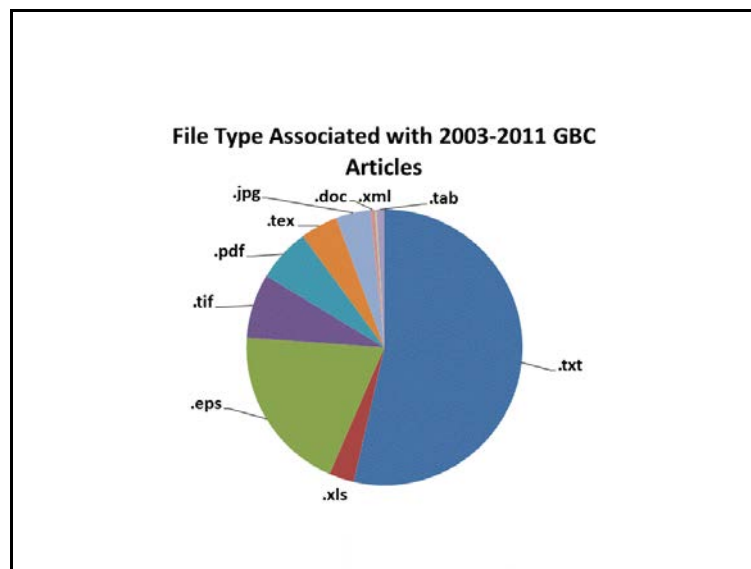


**Figure 3: Submission of Supplemental Materials by Authors of Ocean Science Papers**

Chart prepared by Ed Urban from data available from AGU Web site:

[http://www.agu.org/pubs/authors/manuscript\\_tools/journals/auxiliary\\_material/index.shtml](http://www.agu.org/pubs/authors/manuscript_tools/journals/auxiliary_material/index.shtml)

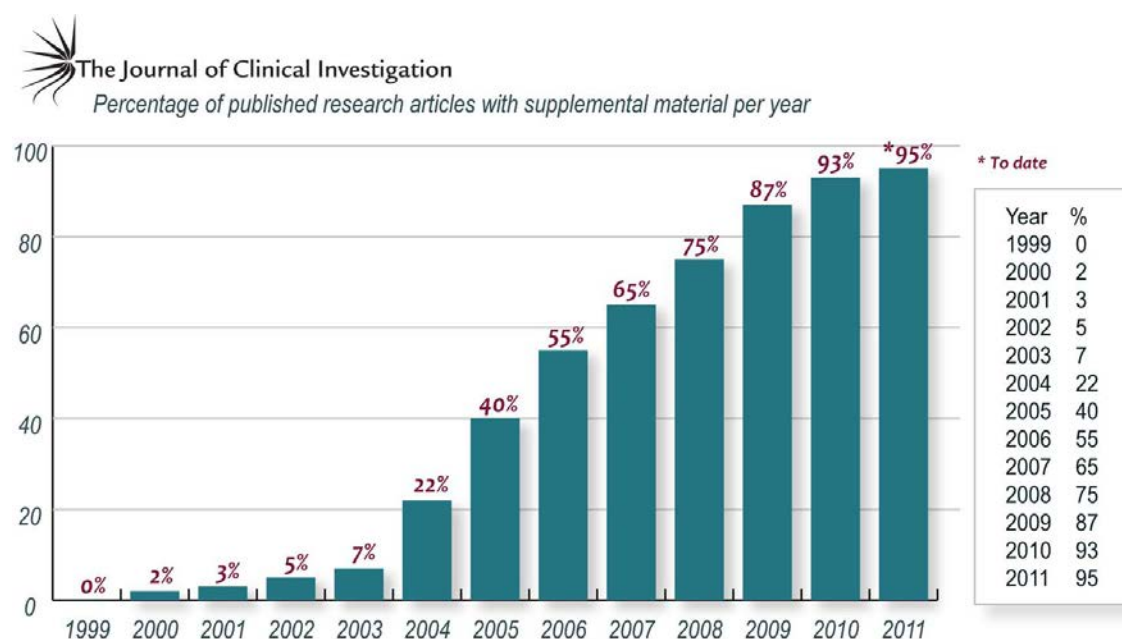
<http://www.agu.org/pubs/journals/>



**Figure 4: File type associated with submissions**

The meeting noted that this concerned publication of data to the AGU data system as “supplemental materials”. So it was not a practice based on institutional policies or practices. The data are open access and not limited to subscribers.

Ms Norton then showed a similar graph (Figure 5) for published articles with supplemental material for the *Journal of Clinical Investigation*.



**Figure 5: Percentage of published research articles with supplemental material per year**

Chart courtesy of Ken Beauchamp, American Society for Clinical Investigation

The meeting noted with concern the contradiction between the statistics provided by Ed Urban and those obtained from the ASCI. The group was unable to explain these differences.

### 3.3 Upstream Provenance: The Fleet as an Observing System

This agenda item was introduced by Mr Bob Arko. He emphasized that the volume and complexity of environmental sensor data from the oceans is growing rapidly, and data are reusable beyond the original mission. U.S. policy on scientific research data increasingly demands open access and reproducible results. Research vessels are both “expeditionary” platforms (supporting dedicated missions) and “observatory” platforms (with a core set of instrument systems “always on”).

Within the framework of the Rolling Deck to Repository (R2R) programme, underway sensor data from U.S. academic research vessels are submitted by vessel technicians, organized with controlled vocabularies, and standard metadata are published with persistent identifiers at the cruise, dataset, and file levels. Quality assessment is performed on selected data types. Data and documentation are ultimately submitted to long-term public archives.

The U.S. NSF Ocean Sciences Division published a revised Data Policy (#11060 of May 2011) that states, under “III. OCE General Data Policy”: *“All underway data collected at sea aboard NSF-supported oceanographic research vessels will be submitted to the appropriate long-term archive through the R2R (Rolling Deck to Repository) program. For these data sets, this relieves the PI and the ship operator of post-cruise data management responsibility for the underway datasets (See Section VI-B for more specific information). For other at-sea, special ops or research-specific datasets collected on a cruise (such as water samples, ROV sampling, dredge hauls, etc). the PI is responsible for collecting and making public the metadata associated with sample collection and the data or results from the research”.*

In addition, under “VI. More Specific Data Submission Guidance”: *“Effective with the release of this policy, the underway data from NSF-funded cruises will be placed in the public domain for unrestricted open access after 60 days of the cruise end date, unless a request for restricted access is submitted through R2R. If such a request is submitted, a proprietary hold period of up to two (2) years will be maintained by R2R with approval of the cognizant NSF Program Officer. Requests for a longer*

proprietary period will be maintained by R2R only with the approval of the cognizant NSF Program Officer. Data will be transmitted to the appropriate National Data Center for dissemination only after the conclusion of the stated proprietary hold”.

The R2R programme plan includes archiving of the original/raw underway sensor data from research vessels at long-term archives (U.S. NOAA Data Centers), and publication of persistent, citable identifiers for such data. These data sets, in turn, may be referenced from the processed/final data sets at downstream systems such as BCO-DMO, and ultimately in journal articles and global syntheses.

The following approach is used:

1. Publish Linked Data resources (URIs that dereference to RDF, in the *rvdata.us* namespace) for all content. This is likely sufficient for most resources such as Cruises, Devices, Formats, Programs, etc.
2. Publish external global identifiers for actual digital objects (Data and Documents):
  - Digital Object Identifiers (DOIs) routinely for data sets
  - Archival Resource Keys (ARKs) selectively for data files, per community demand

As of October 2012, a contract has been established with the CDL/EZID system for unlimited DOIs and ARKs, which includes a programming interface and the option to provide our own IDs (attached to an assigned prefix). Metadata will be published with every DOI that validates to the DataCite Kernel and conforms to ESIP Guidelines. Metadata will be published with every ARK that includes file name, size, date, checksum and parent. DOI/ARK IDs are aligned with internal R2R IDs, but are otherwise entirely opaque.

Guidance is taken from [http://wiki.esipfed.org/index.php/Interagency\\_Data\\_Stewardship/Principles](http://wiki.esipfed.org/index.php/Interagency_Data_Stewardship/Principles)

Two use cases were presented. One is shown below in [Figure 6](#), illustrating the chain of persistent and citable identifiers from original data (DOI#1) to processed data (DOI#2) to journal article (DOI#3).

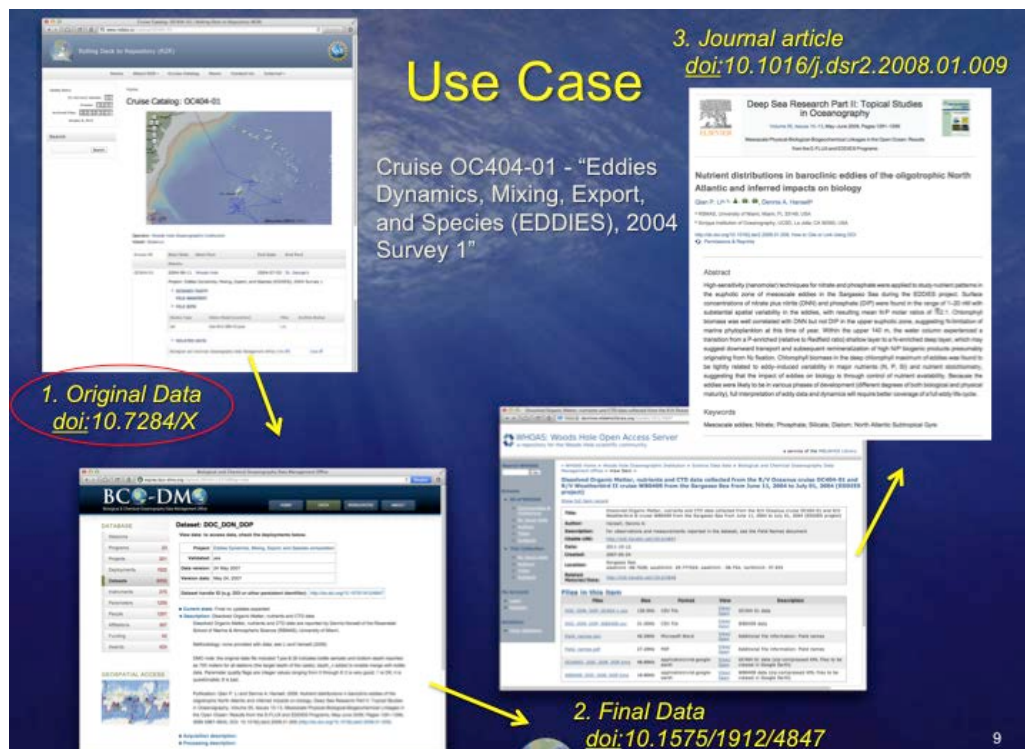


Figure 6: Possible citation pathway for CTD data from EDDIES cruise.

### 3.4 PID interoperability - report from IETF and RDA meetings

This agenda item was introduced by Mr Andrew Maffei. He reported on activities of IETF (Internet Engineering Task Force) and RDA (Research Data Alliance) regarding Permanent Identifiers and Related Topics.

IETF-84 Data Set Identifier Interoperability (DSII) BOF (Birds of a Feather – an ad hoc informal discussion group): Mr Maffei reported that discussion focused on how to achieve interoperability among persistent identifiers for data sets made available on the Internet. The use case of interest was scientific data sets produced by different research teams; other use cases might include media developed by different sources and combined into a common collection. Access policies based on identifiers, discovery, association of meta-data, and data integrity are expected to be later topics, but these will likely be covered in follow-on mailing list discussion. The BOF reviewed existing methods such as DOI, URN, PURL, and then discuss core requirements.

Research Data Alliance (<http://www.rd-alliance.org>) : Mr Maffei informed the group that the vision of the RDA is for researchers around the world sharing and using data without barriers. The purpose of the Research Data Alliance (RDA) is to accelerate international data-driven innovation and discovery by facilitating research data sharing and exchange, use and re-use, standards harmonization, and discoverability. This will be achieved through the development and adoption of infrastructure, policy, practice, standards and other deliverables. The RDA focuses on engineering, not research. The US portion of this initiative is funded by NSF (like EarthCube)

The following workgroups are likely to be proposed at the upcoming “kickoff meeting” in Sweden in March 2012:

- PID Information Types
- ☐ Universal Product Code
- ☐ Data Type Registries
- ☐ Metadata
- ☐ Linking Data and pubs: citations
- ☐ Terminology Harmonization; Semantics
- ☐ Practical Policy
- ☐ Legal
- ☐ Urban Scientist
- ☐ Engaging with Scientists

(please see ☐ <http://d2i.indiana.edu/data2012/wiki/images/c/c2/WG-proposals-All.pdf> for more details)

A slide in the introductory slides at the meeting indicated that one application area of potential interest for RDA involved interactions in marine data because it:

- Requires chemists, physicists, biologists, climatologists
- Leverages marine vocabulary management (ODIP?)
- Test interoperability between oceans, atmospheric and terrestrial data, and climate models
- Test data provenance exchange across the disciplinary boundaries in climate
- Test sharing “data services” rather than “data”- IMOS
- International data registry services

Further information:

- IETF-84 Dataset Identifier Interoperability BOF

July/August 2012 meeting in Vancouver, BC  
<http://www.ietf.org/proceedings/84/dsii.html>

<http://www.ietf.org/proceedings/84/minutes/minutes-84-dsii>

- Research Data Alliance Activities

October 2012 meeting in Washington, DC

[http://d2i.indiana.edu/data2012wiki/images/3/3c/RDA-Washington\\_Opener-2012-10\\_2012Oct01.pdf](http://d2i.indiana.edu/data2012wiki/images/3/3c/RDA-Washington_Opener-2012-10_2012Oct01.pdf)

<http://d2i.indiana.edu/data2012/ResearchDataAlliance>

- Other references

Terminology and Use Cases for Interoperability of identifier Resolution Systems

(<http://tools.ietf.org/html/draft-kahn-dsii-id-res-sys-00>)

## 4. WORK PLAN 2013

### 4.1 Further development of the Cookbook

The meeting considered the “Table of Contents” of the Cookbook, taking into account the presentations and discussions under the previous agenda items and decided on a revised structure.

It was stressed that the cookbook should be usable by the different user groups (**SC**: scientists, **SM**: system manager (including marine librarian and data manager) and **IT**:IT manager). The Cookbook should be structured that allows quick referral to the relevant section(s).

The Group agreed that the document will be drafted further using Google Docs. The proposed deadline for priority items will be 15 November (others are marked \*). A later deadline would be end of January 2012. It was agreed that, as a minimum, the 1-pager should be ready for the AGU Fall Meeting in San Francisco, first week of December.

1. **Executive Summary** (SC, SM, IT)

This should be a 1-page summary describing the requirements and the work flow required to publish scientific data sets. This section will also explain why data publication is important to the different target audiences.

status: not done

author: Adam Leadbetter

2. **Introduction** (SC, SM, IT)

status: mostly done. Explain in more detail the benefits that the research community gets from data publication.

author: Adam

3. **Glossary** (SC, SM, IT)

status: mostly done

author: Cyndy Chandler

4. **Data Publication** (SC, SM, IT)

what is, why and for whom?

status: partially available

author: Adam Leadbetter

- a. What is “Data Publication”?

- b. Why Data Publication?

- c. For whom?  
note: this should include description of the important role of marine librarians (Lisa Raymond, Linda Pikula)

5. **Technology requirements** (SM, IT)

Here we will identify the necessary technology to build the operational system (e-repository software like DSpace, data storage system like ftp site,...). In annexes, some concrete examples can be described. This should also include some cautionary notes, lessons learned (risks of open source, cost of open source (long-term maintenance), security, institutional policies and restrictions, user interface, long-term commitment and stability, disaster recovery, redundancy, estimating storage capacity, estimated load, bandwidth,...)

status: not done

author: Adam Leadbetter, Lisa Raymond, Helge Sagen

- a. DSpace (IT)  
brief description and installation of standard DSpace  
author: Marc Goovaerts (if available), Aditya Kakodkar (IODE Project Office)
- b. Customizing DSpace (adding schemas, adding fields) (IT)  
author: Marc Goovaerts (if available), Aditya Kakodkar (IODE Project Office)
- c. File storage (IT)  
Describing some options  
author: Adam Leadbetter

6. **The data publication process** (SC, SM, IT)

This section will describe the work flow and will include the below (and possibly more) steps.

Insert a diagram in the beginning. Each element will need an easy to understand explanation.

status: mostly done

author: Adam Leadbetter, Lisa Raymond

- a. Assigning a persistent identifier (SM, IT)  
what is a persistent identifier.
  - i. Creating a GUID  
explain what this is. Are there easy-to-use tools for this?
  - ii. Minting a DOI through DataCite  
also mention others – details on each in annex
- b. Metadata (SC, SM)
  - i. Content related fields (SC, SM)
  - ii. Service metadata (SM)
- c. **Data file formats** (SC, IT)  
Include list of recommended formats and those to avoid.  
status: almost done  
author: Adam Leadbetter

- d. **Providing a reference to a published dataset (human readable citation string)** (SC, SM) – normally the SM will create the string, the SC will use the string  
status: done  
author: Adam Leadbetter
- e. **Using data citations** (SC, SM) (\*)  
Examples of publishers that accept or require data citation (table of 20 top journals): what journals require it? Data journals.  
status: not done  
author: Ed Urban
- f. **Use of data citation by institutional management or funding agencies** (SC, SM) (\*)  
How could institutional management use data citation statistics for career advancement? How can funding agencies use these?  
status: not done  
author: tbd

ANNEX : Data publication “best practice” examples

- i. MBL-WHOI
- ii. BODC PDL (including example DataCite xml)
- iii. IODE POD (as a platform for institutions that cannot host their own system)
- iv. R2R

ANNEX : References

The meeting noted that biogeographic/biodiversity data were currently not considered. Mr Helge Sagen provided a link to a document prepared by GBIF in this regard:

ISBN: 87-92020-36-4 - Persistent URI: [http://links.gbif.org/gbif\\_best\\_practice\\_data\\_citation\\_en\\_v1](http://links.gbif.org/gbif_best_practice_data_citation_en_v1)

## 4.2 Further technical developments

The meeting discussed how we might incorporate persistent identifier technology. The discussion evolved to focus on persistent identifiers for people and the creation of an “authority file.” The person ID is important because it would support federated search of resources. Attribution (which would use the person ID) is important but it is not always obvious whose name will be linked to data sets. This could be the journal article author, PI, instrument operator,...

Dr Leadbetter informed the meeting that BODC has a database of people and organizations. In the database they track when individuals change location. For organizations, BODC uses EDMO developed by SeaDataNet.

Mr Pissierssens informed the meeting that IODE has operated OceanExpert since 1997. The database now contains more than 13,000 individual expert records. Work is underway to assign each expert a unique expert ID. There are plans to link OceanExpert to EDMO for organizations. In addition, the OceanDocs e-repository will start using the OceanExpert unique IDs to uniquely identify experts.

Ms Pikula and Ms Raymond informed the meeting about ResearcherID and ORCID, respectively. Ms Pikula informed the meeting about discussions she had with ResearcherID developers: they are not linking into citations for papers not indexed by World of Science. Many scientists in our labs work in

other countries. They have changed focus, specialty and it is difficult to trace their publications. Researcher ID facilitates the tracking of their publications if indexed by WOS. ORCID is an open source system but its continuity is not known and there is fee.

Mr Arko informed the meeting that R2R is mapping from R2R (cruises to persons) to FastLane (funding award database) of U.S. NSF. Ms Chandler added that BCO-DMO does the same.

The meeting recommended that IODE's OceanExpert should add "date of birth" and "dead or alive" as fields.

Mr Arko strongly called for an approach based on federated searching. Mr Pissierssens responded that this approach had been considered by IODE on various occasions, but the conclusion had been that many organizations were not prepared to share directory records that include personal information of their employees.

Ms Norton noted that there are also issues with name changes of people. These need to be managed as well.

It was agreed that Mr Pissierssens would contact Ms Chandler regarding the possible import of expert records of WHOI into OceanExpert, noting that currently only 37 WHOI experts were recorded in OceanExpert.

It was also agreed that Mr Arko, Dr Leadbetter, Ms Chandler, Ms Raymond and Ms Norton would be invited to participate in IODE GE-MIM discussions on further OceanExpert development. In this regard the meeting was informed that the next GE-MIM is planned to be held in Miami, USA between 28 and 31 January 2013. The aforementioned individuals will be invited to participate in the meeting in person or through Webex.

Ms Chandler noted that in the Cookbook we should clarify that the current version of DSpace does not support versioning, but that it will in the future release (possibly as from November 2012).

### **4.3 Promotion and deployment of the use cases in the target communities**

#### Promotion through training

Ms Pikula referred to the IAMSLIC/IODE survey on training needs. The preference for courses was equal between e-repositories and data publication, so the latter can be selected for future courses. A short course on data publication will be organized prior to IODE-XXII and a full course in April/May 2013. Ms Pikula gave a brief overview of topics that will be covered: data publication linking will be emphasized.

The meeting agreed that now that the use cases have been implemented, we need to start training users and promote the establishment of pilot sites.

#### Pilots

Mr Pissierssens suggested starting with existing OceanDocs nodes that already use DSpace and have the basic know-how to install and manage DSpace repositories. He suggested further to identify pilots in both developed and developing regions. He also mentioned IAMSLIC's Aquatic Commons and questioned whether we can accommodate EPrints. He recommended that this is further discussed with IAMSLIC.

Ms Norton suggested we check with IAMSLIC members to check which libraries use DSpace.

The meeting concluded that we should focus for now on DSpace installations. We can then give a demonstration at next year's IAMS LIC Conference and then IAMS LIC can consider whether they wish to join the data publication project.

See also <http://www.dspace.org/whos-using-dspace>

The meeting requested IODE to identify OceanDocs hosts in developing regions. Ukraine, India (NIO) and Kenya (KMFRI) were mentioned. It was also suggested to check whether the IODE-XXII host institution was a DSpace user.

Ms Chandler offered to discuss with the NIO Librarian when she visits NIO in January 2013.

#### Role of researchers

Mr Urban recommended that we should involve scientists in our discussions. He also drew attention to the fact that we have participants only from developed countries. We should involve people from other regions as well. When we meet next time, we should meet at a location where we have started a pilot project. We also need to make sure that the pilot sites have all stakeholder groups involved.

#### Towards a Project

Mr Pissierssens noted that the joint activity between MBLWHOI Library, SCOR and IODE has now reached the point where it is becoming a project. Dr Urban agreed with this statement.

The meeting agreed that IODE, MBLWHOI Library and SCOR should discuss the details of establishing a joint project on data publication.

Dr Urban offered to investigate federal government agency funding opportunities.

Ms Chandler stated that we need to team up with ESIP. We need to remember there is an ESIP working group on data publication ([http://wiki.esipfed.org/index.php/Interagency\\_Data\\_Stewardship/Citations](http://wiki.esipfed.org/index.php/Interagency_Data_Stewardship/Citations)).

## **4.4 Outreach activities**

The meeting then considered outreach activities in 2013-2014.

Mr Pissierssens noted that the Cookbook could be included in the OceanTeacher Digital Library. Courses could also be organized in 2013 on data publication (a first course dealing with e-repositories and data publication is planned in April/May 2013).

Mr Pissierssens suggested that the POGO Conference could be a good venue to promote data publication. It was suggested that the Woods Hole Director, Susan Avery should be approached to make a presentation at the next POGO Session in Cape Town, South Africa in January 2013.

The POGO course in Bermuda could be another suitable opportunity.

The Group recommended that OceanTeacher courses should include a lecture on data publication.

Other opportunities:

- IMDIS (23-25 September, Lucca, Italy) – IODE, Adam Leadbetter, Linda Pikula, Lisa Raymond
- ICES annual science conference (Reykjavik, 23-26 September) – Helge Sagen
- Research Data Alliance (Europe, March 2013). – maybe Andrew Leadbetter
- ESIP (Earth Science Information Partners) (Washington DC, 8-10 January) – Cyndy Chandler
- POGO Conference (January 2013) – Susan Avery (Ms Chandler will check with her)

- IAMSLIC 2013 (Dania Beach, Florida, 20-24 October 2013) – Peter Pissierssens, Lisa Raymond, Linda Pikula

The authors were requested to check whether the paper or reprint can be made available publicly.

## **5. DATE AND PLACE OF NEXT MEETING**

This will depend on the location of the pilot project sites. Timing will probably be around the same time (October 2013).

## **6. CLOSING OF THE MEETING**

Ms Raymond closed the meeting at 13h00 on 10 October. She thanked the participants for their active discussions and welcomed the achievements of this joint initiative.

## **ANNEX I**

### **AGENDA OF THE MEETING**

1. OPENING OF THE MEETING
2. PROGRESS WITH THE USE CASES
  - 2.1 USE CASE 1: MBL WHOI LIBRARY/BCO-DMO
  - 2.2 USE CASE 2: BODC-UHASSELT
  - 2.3 PARTICIPATION IN RCN:OCEANOBSNETWORK
  - 2.4 OUTREACH ACTIVITIES IN 2012
  - 2.5 THE DATA CITATION/DATA PUBLICATION COOKBOOK
3. INVITED SPEAKERS
  - 3.1 CODATA
  - 3.2 NISO WORKING GROUP
  - 3.3 UPSTREAM PROVENANCE: THE FLEET AS AN OBSERVING SYSTEM
  - 3.4 PID INTEROPERABILITY - REPORT FROM IETF AND RDA MEETINGS
4. WORK PLAN 2013
  - 4.1 FURTHER DEVELOPMENT OF THE COOKBOOK
  - 4.2 FURTHER TECHNICAL DEVELOPMENTS
  - 4.3 PROMOTION AND DEPLOYMENT OF THE USE CASES IN THE TARGET COMMUNITIES
  - 4.4 OUTREACH ACTIVITIES
5. DATE AND PLACE OF NEXT MEETING
6. CLOSING OF THE MEETING



## ANNEX II

### LIST OF PARTICIPANTS

Mr Robert ARKO  
Lead Programmer/Analyst  
Lamont-Doherty Earth Observatory  
61 Route 9W  
Palisades NY 10964  
United States  
Tel: 1-845-365-8752  
Email: [arko@ldeo.columbia.edu](mailto:arko@ldeo.columbia.edu)

Ms Cyndy CHANDLER  
Informatics Specialist  
MS 36, Woods Hole, MA, 02543  
United States  
Tel: 1-508-289-2765  
Fax: 1-508-289-2009  
Email: [cchandler@whoi.edu](mailto:cchandler@whoi.edu)

Dr Adam LEADBETTER  
Data Scientist  
British Oceanographic Data Centre  
Joseph Proudman Building  
6 Brownlow Street  
Liverpool  
L3 5DA  
United Kingdom  
Tel: +44 (0)151 795 4878  
Email: [alead@bodc.ac.uk](mailto:alead@bodc.ac.uk)

Mr Andrew MAFFEI  
Senior Information Systems Specialist  
Woods Hole Oceanographic Institution  
Mailstop 44  
Woods Hole Massachusetts 02543  
United States  
Tel: 1-508-289-2764  
Email: [amaffei@whoi.edu](mailto:amaffei@whoi.edu)

Ms Cathy NORTON  
Senior Library Scholar  
MBLWHOI Library  
7 MBL ST  
Woods Hole MA 02543  
United States  
Tel: 01 508 560 0617  
Email: [cnorton@mbi.edu](mailto:cnorton@mbi.edu)

Ms Linda PIKULA  
Regional Librarian

NOAA Central and Regional Libraries  
4301 Rickenbacker Causeway  
Miami Florida 33149  
United States  
Tel: 305-361-4429  
Fax: 305-361-4552  
Email: [linda.pikula@noaa.gov](mailto:linda.pikula@noaa.gov)

Mr Peter PISSIERSSENS  
Head, IOC Project Office for IODE,  
Oostende, Belgium  
UNESCO/IOC Project Office for  
IODC  
Wandelaarkaai 7 - Pakhuis 61  
B-8400 Oostende  
Belgium  
Tel: +32-59-340158  
Fax: +32-59-79 5220  
Email: [p.pissierssens@unesco.org](mailto:p.pissierssens@unesco.org)

Ms Lisa RAYMOND  
Co-Director MBLWHOI Library  
Woods Hole Oceanographic Institution  
Woods Hole MA 02543  
United States  
Tel: 1 508 289 3557  
Fax: 1 508 457 2156  
Email: [lraymond@whoi.edu](mailto:lraymond@whoi.edu)

Mr Helge SAGEN  
Head of Norwegian Marine Datacentre  
(NODC)  
Institute of Marine Research, Bergen  
Nordnesgaten 33  
Postboks 1870 Nordnes  
5817 Bergen  
Norway  
Tel: +4755238447  
Fax: +4755238531  
Email: [Helge.Sagen@imr.no](mailto:Helge.Sagen@imr.no)

Dr Ed URBAN  
Executive Director  
Scientific Committee on Oceanic  
Research  
Robinson Hall  
University of Delaware  
Newark DE 19716

United States  
Tel: 302-831-7011 / / 1  
Fax: 302-831-7012 / / 1  
Email: [ed.urban@scor-int.org](mailto:ed.urban@scor-int.org)





## **Intergovernmental Oceanographic Commission (IOC)**

United Nations Educational, Scientific and Cultural Organization (UNESCO)

1, rue Miollis, 75732 Paris Cedex 15, France

Tel: +33 1 45 68 39 83

Fax: +33 1 45 68 58 12

<http://ioc.unesco.org>

IOC Project Office for IODE

Wandelaarkaai 7

8400 Oostende, Belgium

Tel: +32 59 34 21 34

Fax: + 32 59 34 01 52

<http://www.iode.org>