

# A Woods Hole Data Repository: Addressing the Issues of Provenance, Attribution, Citation, and Accessibility

## Executive Summary

THE MBLWHOI Library has been working with stakeholders in the management and publication of data with support from the George Frederick Jewett Foundation. Along with the Library Director and the Data Librarian two informaticians have been supported, Andy Maffei and Dr. Holly Miller to bring the community together in approaching standards for publication. In April a Data Attribution and Provenance for Published Datasets Workshop was held at the National Academy of Sciences, Jonson Center 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 new requirements that authors make the data underlying the figures, tables and text of submitted manuscripts available for readers and other interested parties. 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 (henceforth designated "backbone data") to be accessible it needs to be (1) discoverable, (2) citeable and (3) available on the Internet. Resources, standards, and workflows must be defined to support the publisher and funding agency mandates. For the backbone data to be discoverable, appropriate metadata, defined using metadata standards utilizing community accepted ontologies, must be associated with the data file. Backbone data will be made citeable by the assignment of a persistent identifier and provenance metadata and attribution. The availability of the backbone data will be assured by submission to a data repository that has stability and permanence.

The Use Case chosen for the workshop describes a scientist who wants to publish the data associated with the article he is submitting, "Acoustic properties of *Salpa thompsoni*" to a journal. This backbone data for this article includes images and datasets used to generate figures and tables in the article. Minimum metadata fields based on Dublin Core Schema with the addition of some Darwin Core fields for describing the data were decided upon during the workshop and subsequent discussions between Peter Wiebe and MBLWHOI Library staff. Digital Object Identifiers (DOI's) will be assigned to the datasets to make the data easy to cite and retrieve. The MBLWHOI Library's DSpace installation, the Woods Hole Open Access Server (WHOAS), was chosen to be the data repository for the Use Case data. WHOAS had already been accepting data, but the new challenge is to integrate additional metadata fields and to develop a workflow that would facilitate author submission to journals and incorporate DOI's.

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 identified include limited incentives for researchers to expend extra resources to publish their data, fears of data depositors of theft and loss of control over their data. Cultural challenges will be overcome as funding agencies pressure scientist 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 a second workshop in March 2010 progress will be shared with a panel of marine and oceanographic experts for feedback and to work through challenges encountered. New ways of presenting data to scientists and the public will also be explored. The challenges associated with storing, organizing and making data accessible are many however the anticipated benefits to the scientific and global community compel us to continue to meet those challenges.

### **Introduction**

Publishers and funding agencies have begun to require authors to make the data that supports the figures, tables, and text in their scientific papers accessible. This will require a process that effectively captures, publishes, preserves, and tracks relevant metadata and preserves provenance and attribution. Four criteria must be met in order for this to occur. The data must be:

- 1) Discoverable: Appropriate metadata, using community-accepted metadata standards must be associated with data.
- 2) Citable: A persistent identifier and provenance metadata and attribution must be assigned to the data file.
- 3) Available on the Internet: A data repository that is both stable and permanent must be created so that the availability of data will be assured.
- 4) Reusable: Sufficient metadata, provenance, and attribution information is required to enable reuse of the data

The challenges associated with meeting these criteria fall into three categories:

- 1) Technical: As data acquisition becomes automated, the rate and volume of data has increased dramatically. Prior to the advent of digitization, a great deal of supportive data simply wasn't published, or if it was, it tended to be in smaller amounts and/or as a tangible item, such as a table in a paper. Metadata has been insufficient, non-standard, or lacked a persistent identifier. In addition, even when metadata has existed, it has not been automated at the same rate as the data itself.
- 2) Cultural: Until recently, researchers have had little or no motivation for making their source data available. Moreover, many researchers feel proprietary about the data they have collected, and are concerned that it might be stolen, misused, or used without proper attribution.
- 3) Common/Usual: There is a wide range of issues that have added to the difficulty of creating an effective data attribution and provenance system. They include:
  - Lack of resources
  - Lack of funding
  - Lack of time
  - Lack of expertise
  - Lack of protocols

## The Workshop

The “Data Attribution and Provenance for Published Datasets Workshop,” sponsored by the MBLWHOI Library and funded by the George Frederick Jewett Foundation, was held on April 9 – 10, 2009 in Woods Hole, MA. The workshop was designed to define the challenges inherent in creating an effective system for publishing source data associated with scientific journal articles and to suggest best practices for standards and workflow practices to address these challenges. The workshop (see *Appendix I: Workshop Agenda*) focused on establishing best practice guidelines for data attribution and provenance. The challenges and scope of the problem were illuminated with several presentations by participants (see *Appendix II: Workshop Participants*). Topics included: status of data at the Data Library and Archives at the MBLWHOI Library; persistent archives; Rensselaer Polytechnic Institute (RPI) data projects; data standards; and the value of data sharing. Preceding and during the workshop, articles, websites, and other information resources (*Appendix III – Recommended Reading List*) were shared by the participants and a wiki ([http://tw.rpi.edu/portal/Jewett\\_Meeting\\_at\\_MBL](http://tw.rpi.edu/portal/Jewett_Meeting_at_MBL)) was created to store the additional meeting notes, the PowerPoint presentations, and other information among the participants and with the larger community.

The following themes were used to direct and guide the workshop sessions:

- Focus on the data behind published journal articles, also referred to as “backbone data.
- Examine metadata, provenance, and attribution information, with a goal of determining how much is enough
- Best practices for data citation
- Where and how should data and metadata should be stored
- Required/recommended metadata
- Reuse and misuse of data

In order to meet these goals an actual Use Case, an article written for publication by Dr. Peter Wiebe, was used to test the process of generating, assigning, depositing, and storing published backbone data with appropriate metadata. Finally, specific solutions were proposed, presented, and action items identified.

## The Use Case

The use case chosen for the workshop describes an actual scientist (Peter Wiebe) who wanted to publish the backbone data associated with the article he was submitting for publication. The title of the article was, “Acoustic Properties of *Salpa thomsoni*.” His data included images and datasets used to generate figures and tables in the article.

### Focus I: Data Provenance

In order to provide reliable and accurate data provenance, some fundamental questions must be addressed:

- What is the fundamental definition of source or “backbone” data?
- If backbone data is considered a subset of raw data, how much of the raw data must be archived to support the backbone data?

- How will the raw source data be broken and processed into subsets?
- How far back do you go when storing data?
- What is the authority chain? Who owns the data, and who determines how and by whom it is used?
- What is the best way to make ownership of the data clear to outside users?
- What is the best way to make sure that accurate citations are used and maintained as data is used again and again?
- Where should the data, metadata, source data, and attribution streams be stored?
- What is the best way to encourage/promote the use of metadata?

## **Focus II: Metadata Generation and Formatting**

Metadata is used for storage purposes, but also as a way to communicate the origin, authority, methods used to process the data, and how potential users can use it. Enough information is needed to be able to use, recompile, add, and merge data. In order to be useful, metadata needs to:

- Use a standard format
- Be as rich as possible – give as much description as needed for discovery and reuse
- Presented in a format that is easy to understand
- Use controlled vocabulary or ontology
- Be machine readable

It is important to work with scientists to create the metadata template so it can be integrated into the researcher's workflow. Metadata fields that should be included in the template were:

- Latitude/longitude
- Data source
- Link outs: every piece of metadata used must be accurately linked to the figures and tables it supports. Moreover, these links must persist over time, which means that a reliable, long-term infrastructure needs to be in place.
- Process descriptions written in natural language so that no meaning is lost when converting to template/controlled vocabulary style data
- Cruise ID or cruise number – USGS uses a field activity number.

It was also noted that scientific domains work with widely differing data forms and file types and therefore may require different metadata templates. To insure that all data types were accommodated, it was proposed that a taxonomy of data types be created.

## **Focus III: Attribution**

The major issue surrounding attribution is the number and kinds of contributors to the backbone data. This raised questions about the number and kinds of fields needed to produce metadata that is genuinely useful.

A modified version of Dublin Core and Darwin Core metadata, that would accommodate descriptions of the backbone data, was proposed. This led to a discussion of what new fields should be added to the metadata, at what level DOIs should be granted, and what new tools would be needed to extract creator data.

In order to produce accurate metadata attribution fields the attribution stream must be clear. Four kinds of creators were identified for attribution:

- Data creator: There was some discussion about the definition of a data creator. If all data is being used, then the original data creator/generator should be credited as co-author on the paper. If only a small portion of the data is used, then the data creator/generator and author of the paper need to negotiate. “Rules of the road” need to be established at the repository level to insure that those accessing the data adhere to this policy.
- Journal article (tables, figures) creator
- Metadata creator
- Principle investigator – the main scientist responsible for designing the sampling protocols for the data gathering project

Based on these four kinds of creators, the following attribution element names were suggested:

- dc.contributor – should be the same for all figures/tables in the publication
- dc.creator
- wc.dataattribution (note that the “wc” prefix denotes the Woods Hole Core)

The question was then raised about when and at what level Digital Object Identifiers (DOI) should be granted. It was proposed that a DOI be granted for each dataset. This would allow the DOI to serve as a snapshot of the overall data in the set. A researcher could then access and download the whole dataset as a tar bundle containing images and grids. It would be up to the user to figure out which files to use and the systems needed to read them.

#### **Focus IV: Where to store data**

The main focus of this section of the workshop was where backbone data and metadata associated with publications should be stored. The following questions guided the discussion:

- Should backbone data and metadata be stored in the same place?
- Should the data and metadata be stored in a library?
- What role should libraries take in storing and providing access to data?
- Do librarians have the expertise necessary to manage QA/QC of data?
- What are the guidelines for monitoring and controlling access? DOIs do not allow the names of those who access the data to be tracked.
- Do the same professional standards that apply to the reuse of text apply to the reuse of data?

The MBWHOI Library is a trusted information source that will be here long after grants have ended and researchers move on to other projects. The Library has long been preserving and making available the output of Woods Hole science in text form, it is a logical step to now link the published articles to the data that supports them.

It was proposed that MBLWHOI Library set up a prototype using the Wiebe Use Case, in order to determine the requirements for developing a workflow, processing the data and metadata and making it available. Other data stores (such as BCO-DMO) perform quality assurance (QA) and

quality control (QC) on the data submitted. However, currently the library does not have the staff to perform more than minimal review of the data submitted. Future discussions and action steps will further define the amount of QA/QC needed and infrastructure for implementing it.

### **Progress to Date**

It was determined during the Data Attribution and Provenance Workshop that the Library would to accept the data that supports figures and tables from an article for publication into the Woods Hole Open Access Server (WHOAS) with the Wiebe Case Study chosen to be the test case.

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 review previous mapping. Workshop participants focused on the Dublin Core schema, but also invented some fields, referring to them as Woods Hole Core. To move the project along, 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 follow up meeting with the author prompted the addition of some Darwin Core fields, dwc.scientificName and dwc.genus. More Darwin Core fields can be added, if requested. 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.

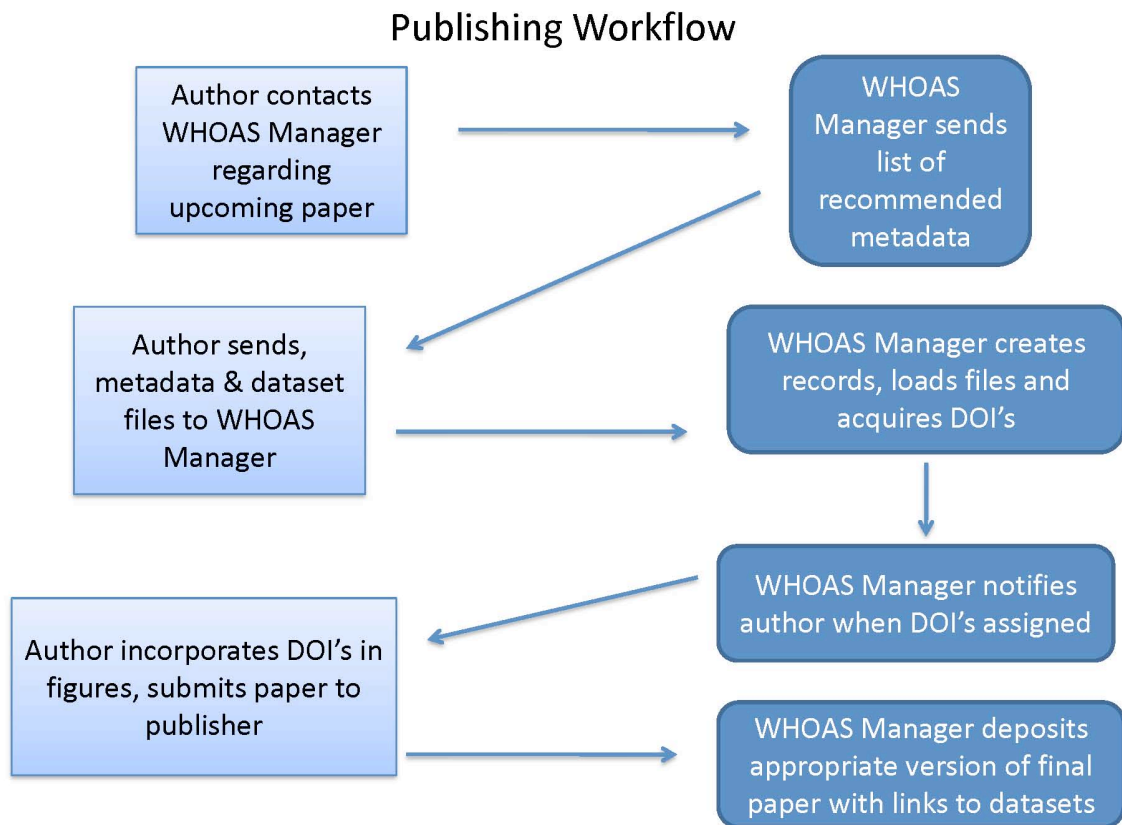


Figure 0 Publishing workflow for processing and storing backbone data associated with a scientific publication.

The next task was to develop a scalable workflow (*Figure 1*) that incorporates the assignment of DOI's to each dataset for figures and tables. This workflow must be timed to coordinate with the publication process in such a way that DOI's 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 assures DOI's are assigned in time to be included in the final publication. It could result in assignment of DOI's 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 DOI's to figures and tables that don't get published. The author can decide if the unpublished datasets should have links in 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 DOI's 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. He 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. The other reason to expect more progress with DSpace programming is the recent merger of DSpace and Fedora and the infusion of funds through a new membership program. 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 Dr. Wiebe over the summer and through the fall, requesting the datasets and some additional metadata. Time got away from him and the paper was accepted for publication and the final draft sent out before the datasets could be submitted and DOI's acquired 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.

Other authors have been identified and contacted about testing the workflow and one MBL researcher has committed to the project. We believe that we contacted him early enough in the data collecting and publishing process that we might mitigate the problems associated with trying to go back and put together the metadata and organize the datasets.

Library staff made a brief oral presentation at the International Association of Aquatic and Marine Science Libraries and Information Centers (IAMSLIC) Conference in Brugge Belgium about this data repository and followed up with a powerpoint presentation via video conference to a group assembled for an IODE OceanData Portal meeting in Oostende Belgium. We expect

further collaboration with IODE and look forward to a joint meeting with the SCOR/IODE Data Publishing group.

### **Future Directions**

The library staff will continue to recruit scientists from MBL and WHOI to submit backbone data with associated metadata to further refine the workflow and processes. The next use case will be a manuscript in preparation by Peter Smith, an electrophysiologist and director of the Biocurrents Research Center at MBL.

A second workshop is planned for the end of March 2010. This meeting is expected to include participants from the SCOR/IODE Workshop on Data Publishing (2008). This meeting will bring together an international panel of experts working on marine and oceanographic data. Workshop topics will include: standards for metadata and citation; case studies – challenges and success; and sustainability.

The library is also embarking on an experimental project exploring ways to visually represent data. Combining aesthetic, empirical, and mathematical disciplines, data visualization harnesses multiple skill sets to add meaning and clarity to complex numerical information through graphical means. In addition to being beautiful, data visualizations are often the most effective way to communicate and analyze large, multifaceted data sets. The Library will explore data visualization in three dimensions by building a Spatial Kinetic Display that allows viewers to investigate biodiversity data.

## Appendix I - Meeting Agenda

Thursday, April 9th

4:00-5:00 pm Keynote by Deborah McGuinness (RPI)

5:00-5:30 pm Challenges by Cyndy Chandler (WHOI)

5:30-6:00 pm Goals - discussion Andy Maffei (WHOI)/Cathy Norton (MBLWHOI Library)

- focus is only on data behind a published journal article
- examine attribution stream for this data, how is it cited?
- examine where do you store the metadata about this data?
- where do you store the data?
- what metadata is required around the metadata?

Friday, April 10th

8:30-9:00 am Data Library by Lisa Raymond (MBLWHOI Library)

9:00-9:30 am Persistent Archives: Long Term Sustainability of data based on policy and data virtualization by Arcot Rajasekar (UNC)

9:30-10:00 am NSF Office of CyberInfrastructure : What Are We Thinking About Data by Jennifer Schopf (NSF)

10:00-10:30 am Break

10:30-Noon Practicum - Use Cases

Noon Lunch - Jonsson Center/ Main House

1:00-1:30 pm Data Standards, Better Practices: US and others by Peter Fox (RPI)

1:30-3:00 pm - Use cases continued - followed by breakouts if necessary

3:00-3:30 pm Break

3:30-6:00 pm Consensus on Best Practices.... and work on white paper resulting from discussions.

## Appendix II - List of Participants

Cyndy Chandler  
Data Manager, Biological and Chemical  
Oceanography Data Management Office  
(BCO-DMO)  
Woods Hole Oceanographic Institution  
Shiverick 102E, MS#36  
Woods Hole, MA 02543  
United States  
Tel: 508-289-2765  
email: [cchandler@whoi.edu](mailto:cchandler@whoi.edu)

Li Ding  
Research Scientist  
Rensselaer Polytechnic Institute  
Lally 2nd Fl  
110 8th Street  
Troy, NY 12180  
United States  
Tel: 518-276-4426  
email: [dingl@cs.rpi.edu](mailto:dingl@cs.rpi.edu)

Vicki Ferrini  
Associate Research Scientist  
Lamont-Doherty Earth Observatory  
208A Oceanography  
61 Route 9W - PO Box 1000  
Palisades, NY 10964-8000  
United States  
Tel: (845) 365-8339  
email: [ferrini@ldeo.columbia.edu](mailto:ferrini@ldeo.columbia.edu)

Peter Fox  
Professor and Tetherless World Constellation  
Chair  
Rensselaer Polytechnic Institute  
WINSLOW BUILDING 2nd fl  
110 8th Street  
Troy, NY 12180  
United States  
Tel: 518-276-4862  
email: [pfox@cs.rpi.edu](mailto:pfox@cs.rpi.edu)

Art Gaylord  
CIS Director  
Woods Hole Oceanographic Institution  
Clark 140, MS#46  
Woods Hole, MA 02543  
United States  
Tel: 508-289-3329  
email: [agaylord@whoi.edu](mailto:agaylord@whoi.edu)

Anthony Goddard  
Systems Administrator  
MBLWHOI Library  
7 MBL St.  
Woods Hole, MA 02543  
United States  
Tel: 508-289-7649  
email: [agoddard@mbl.edu](mailto:agoddard@mbl.edu)

Robert Groman  
Data Manager, Biological and Chemical  
Oceanography Data Management Office  
(BCO-DMO)  
Woods Hole Oceanographic Institution  
Shiverick 102E, MS#36  
Woods Hole, MA 02543  
United States  
Tel: 508-289-2409  
email: [rgroman@whoi.edu](mailto:rgroman@whoi.edu)

Kerstin Lehnert  
Administrative Director of Research  
Lamont-Doherty Earth Observatory  
213 Monell  
61 Route 9W - PO Box 1000  
Palisade, NY 10964-8000  
United States  
Tel: 845-365-8506  
email: [lehnert@ldeo.columbia.edu](mailto:lehnert@ldeo.columbia.edu)

Andy Maffei  
Senior Information Systems Specialist  
Woods Hole Oceanographic Institution  
Clark 151A, MS#46  
Woods Hole, MA 02543  
United States  
Tel: 508-289-2764  
email: [amaffei@whoi.edu](mailto:amaffei@whoi.edu)

Deborah L. McGuinness  
Sr. Constellation Prof. of Tetherless World  
Res. Constellation  
Rensselaer Polytechnic Institute  
WINSLOW BUILDING 2nd fl  
110 8th Street  
Troy, NY 12180  
United States  
Tel: 518-276-4404  
email: [d1m@cs.rpi.edu](mailto:d1m@cs.rpi.edu)

Greg Miller  
Physical Scientist  
Web Master, Woods Hole Field Center  
United States Geological Survey  
Woods Hole Science Center  
384 Woods Hole Road  
Quissett Campus  
Woods Hole, MA 02543-1598  
United States  
Tel: 508-457-2293  
Email: [gmliller@usgs.gov](mailto:gmliller@usgs.gov)

Holly Miller  
Project Leader, Biology of Aging  
MBLWHOI Library  
7 MBL St.  
Woods Hole, MA 02543  
United States  
Tel: 508-289-7632  
email: [hmliller@mbl.edu](mailto:hmliller@mbl.edu)

Stephen Miller  
Head, Geological Data Center  
Scripps Institution of Oceanography, UCSD  
9500 Gilman Drive  
La Jolla, CA 92093  
Mail Code: 0220  
United States  
email: [spmiller@ucsd.edu](mailto:spmiller@ucsd.edu)

Tom Moritz  
Principal, Tom Moritz Consultancy  
Conservation Commons Steering Committee  
at IUCN - World Conservation Union  
Tel: 310-963-0199  
Email: [tom.moritz@gmail.com](mailto:tom.moritz@gmail.com)

Art Newhall  
Information Systems Specialist  
Woods Hole Oceanographic Institution  
Bigelow 210-B, MS#11  
Woods Hole, MA. 02543  
United States  
Tel: 508-289-3317  
email: [anewhall@whoi.edu](mailto:anewhall@whoi.edu)

Cathy Norton  
Library Director  
MBLWHOI Library  
7 MBL St.  
Woods Hole, MA 02543  
United States  
Tel: 508-289-7341  
email: [cnorton@mbl.edu](mailto:cnorton@mbl.edu)

Alice Orton  
Data Cataloger  
United States Geological Survey  
Woods Hole Science Center  
384 Woods Hole Road  
Quissett Campus  
Woods Hole, MA 02543-1598  
United States  
Tel: 508-457-2335  
Email: [aorton@usgs.gov](mailto:aorton@usgs.gov)

Arcot Rajasekar  
Director of Research and Technologies in the  
Data Intensive Cyber Environments Center  
(DICE Center), Professor in the School of  
Library and Information Sciences at the  
University of North Carolina at Chapel Hill,  
and a Chief Scientist at the Renaissance  
Computing Institute (RENCI)  
University of North Carolina at Chapel Hill  
Manning Hall Room 202  
RENCI Room 556A  
Chapel Hill, NC 27599-3360  
United States  
Tel: 858/534-8378  
sekar@diceresearch.org

Lisa Raymond  
Assistant Library Director, Manager Data  
Library and Archives  
MBLWHOI Library  
Woods Hole Oceanographic Institution  
McLean 114, MS#08  
Woods Hole, MA. 02543  
United States  
Tel: 508-289-3557  
email: lraymond@whoi.edu

Ryan Schenk  
Web Application Architect  
MBLWHOI Library  
7 MBL St.  
Woods Hole, MA 02543  
United States  
Tel: 508-289-7548  
email: [rschenk@mbl.edu](mailto:rschenk@mbl.edu)

Jennifer Schopf  
Program Director, Office of  
Cyberinfrastructure  
National Science Foundation  
4201 Wilson Boulevard, Room 1160 N  
Arlington, Virginia 22230  
United States  
Tel: 703-292-4770  
email: [jschopf@nsf.gov](mailto:jschopf@nsf.gov)

Edward Urban  
Executive Director  
Scientific Committee on Oceanic Research  
(SCOR) Secretariat  
College of Earth, Ocean, and Environment  
Robinson Hall  
University of Delaware  
Newark, DE 19716  
United States  
Tel: 302-831-7011  
email: [ed.urban@scor-int.org](mailto:ed.urban@scor-int.org)

Patrick C. West  
Systems Engineer, Sr.  
Rensselaer Polytechnic Institute  
SC 1st Fl  
110 8th Street  
Troy, NY 12180  
United States  
email: [westp@rpi.edu](mailto:westp@rpi.edu)

Peter Wiebe  
Scientist Emeritus  
Woods Hole Oceanographic Institution  
Redfield 2-26, MS#33  
Woods Hole, MA 02543  
United States  
Tel: 508-289-2313  
email: [pwiebe@whoi.edu](mailto:pwiebe@whoi.edu)

Stephan T. Zednik  
Systems Engineer  
Rensselaer Polytechnic Institute  
Tr 3rd Fl  
110 8th Street  
Troy, NY 12180  
United States  
email: [zednis@rpi.edu](mailto:zednis@rpi.edu)

### **Appendix III - Recommended Reading List**

National Science and Technology Council Releases Strategy for Digital Scientific Data A view down the middle of a boron nitride nanotube. The National Science and Technology Council (NSTC) released a report describing a strategy to promote preservation and access to digital scientific data. The report, Harnessing the Power of Digital Data for Science and Society, was produced by the NSTC's Committee on Science under the auspices of the Office of Science and Technology Policy (OSTP) in the Executive Office of the President. The open and timely publication of digital scientific data called for in the report will ... More at [http://www.nsf.gov/news/news\\_summ.jsp?cntn\\_id=114448&govDel=USNSF](http://www.nsf.gov/news/news_summ.jsp?cntn_id=114448&govDel=USNSF) 51

Survey of data provenance techniques. Technical Report IUB-CS-TR618  
<http://www.cs.usask.ca/faculty/sal426/Provenance/docs/Literature%20Review/TR618.pdf>

ICSU Ad Hoc Strategic Committee on Information and Data  
[http://www.icsu.org/Gestion/img/ICSU\\_DOC\\_DOWNLOAD/2123\\_DD\\_FILE\\_SCID\\_Report.pdf](http://www.icsu.org/Gestion/img/ICSU_DOC_DOWNLOAD/2123_DD_FILE_SCID_Report.pdf)

Sudha Ram, Jun Liu. Understanding the Semantics of Data Provenance to Support Active Conceptual Modeling <http://en.scientificcommons.org/41046974>

Fox, McGuinness, Pinheiro da Silva. Knowledge Provenance in Virtual Observatories: Applications to Image Data Pipelines, 2008.  
[http://data.semanticweb.org/conference/iswc/2008/paper/poster\\_demo/70/html](http://data.semanticweb.org/conference/iswc/2008/paper/poster_demo/70/html)

Pinheiro da Silva, McGuinness, McCool. Knowledge Provenance Infrastructure.  
<http://en.scientificcommons.org/685801>

Clifford Lynch. The Shape of the Scientific Article in the Developing Cyberinfrastructure,” CTWatch Quarterly (August 2007) <http://www.ctwatch.org/quarterly/articles/2007/08/the-shape-of-the-scientific-article-in-the-developing-cyberinfrastructure/>

Skills, Role & Career Structure of Data Scientists & Curators: Assessment of Current Practices & Future Needs. JISC Report 2008.  
<http://www.jisc.ac.uk/publications/publications/dataskillscareersfinalreport.aspx>

Baker, Barton, Peterson, Fox. Informatics and the 2007-2008 Electronic Geophysical Year. EOS, Transactions, American Geophysical Union 89(48) 2008.  
<http://www.agu.org/pubs/crossref/2008/2008EO480001.shtml> (subscription)

Gomes, Graybeal and O'Reilly. Data Management Issues in Operational Ocean Observatories. Sea Technology 48(5) p.17-20, 2007 <http://www.highbeam.com/doc/1P3-1284688471.html> (subscription)

Altman and King. A Proposed Standard for the Scholarly Citation of Quantitative Data  
<http://gking.harvard.edu/files/cite.pdf>

Trustworthy Repositories Audit & Certification: Criteria and Checklist  
<http://www.crl.edu/PDF/trac.pdf>

SCOR/IODE Workshop on Data Publishing, Oostende, Belgium, 17-19 June 2008. UNESCO, 2008. IOC Workshop Report No. 207.

[http://www.iode.org/index.php?option=com\\_oe&task=viewDocumentRecord&docID=2457](http://www.iode.org/index.php?option=com_oe&task=viewDocumentRecord&docID=2457)

Standards for DATA

A Proposed Standard for the Scholarly Citation of Quantitative Data

<http://gking.harvard.edu/files/cite.pdf>

ISO 8000 under development

[http://www.iso.org/iso/iso\\_catalogue/catalogue\\_tc/catalogue\\_detail.htm?csnumber=50801](http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=50801)

ISO 19115

[http://www.iso.org/iso/iso\\_catalogue/catalogue\\_tc/catalogue\\_detail.htm?csnumber=26020](http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=26020)

ISO 19115:2003 defines the conceptual model required for describing geographic information and services. It provides information about the identification, the extent, the quality, the spatial and temporal schema, spatial reference, and distribution of digital geographic data