United Nations
Educational, Scientific and
Cultural Organization

Intergovernmental
Oceanographic
Commission

Workshop Report No. 207

SCOR/IODE WORKSHOP ON DATA PUBLISHING

IOC Project Office for IODE, Oostende, Belgium, 17-18 June 2008

International Council for Science
SCOR
Scientific Committee on Oceanic Research

UNESCO 2008

IOC Workshop Report No. 207
Oostende,
English only

Workshop Participants (left to right: M. Diepenbroek, P. Pissierssens, P. Simpson, R. Pepe, P. Wiebe, M. Costello, F. Werner, E. Urban, C. Lenhardt, R. Lowry, C. Emerson) (F. Hernandes and J. Verheggen are not included in the picture)

For bibliographic purposes this document should be cited as follows:

SCOR/IODE Workshop on Data Publishing, Oostende, Belgium, 17-19 June 2008.
Paris, UNESCO, 23pp. 2008. (IOC Workshop Report No. 207) (English)

TABLE OF CONTENTS

ANNEXES

1. OPENING OF THE WORKSHOP

Dr Ed Urban, Executive Director of the Scientific Committee on Oceanic Research (SCOR) and Mr Peter Pissierssens, IODE Programme Coordinator and Head of the IOC Project Office for IODE welcomed the participants to the Meeting.

Dr Urban recalled that several meetings of SCOR projects that discussed data management had raised the issues that (1) data and metadata data are not getting into appropriate databases and (2) more incentives, such as credit for publishing data, must be implemented to induce scientists to make their data more available for other scientists. As IODE had similar concerns it was decided to organize this workshop as a joint SCOR/IODE event and to study and further develop solutions to the stated problems. In this regard it was noted that SCOR has wide and easy access to the ocean research community and IODE represents 76 National Oceanographic Data Centres (NODCs). The participants in the Workshop represent a wide range of expertise including information and library sciences, data management, ocean science, ocean modelling, and both primary and secondary publishing.

Mr Pissierssens explained that the issue of data citation and data publication had been discussed within the IODE community about 5 years ago when the IODE Committee recognized that substantial data volumes were currently not being submitted to the NODCs. It had been observed that research scientists perceive that the additional work involved in submitting data sets to NODCs outweighs the advantages of having data archived in these data centres. The ability to "publish" data sets as unique objects and their citation by other researchers could be the missing incentive to improve data flow to NODCs.

Dr Urban summarized the objectives of the Workshop as follows:

(i) describe current status of data citation and publication in oceanography;
(ii) identify problem areas in data citation and publication in oceanography;
(iii) identify interoperability issues of currently used data citation and publication in oceanography; and
(iv) formulate suggestions to address problem areas in data citation and publication in oceanography.

The agenda of the meeting is attached as Annex I. The list of participants is attached as Annex II. A list of definitions of terminology used in this document is attached as Annex III and a list of acronyms used is attached as Annex IV.

2. PROBLEM STATEMENT

There are substantial amounts of data collected by ocean scientists that are not deposited in existing data centres or archive sites. Unfortunately, the longer data sets remain isolated within investigators' laboratories, the more likely it will be that the data will be lost to the community. This is a widely recognized problem, but one that has not seen concerted community action to find solutions, in spite of the fact that many funding agencies are now mandating that data from publicly funded research be made freely available (Costello and Vanden Berghe 2006)[1].

The rapid evolution of the high-speed Internet and the availability of large digital storage capacities have enabled the transfer and storage of comprehensive data sets. Tools for integration and management of disparate data sets are rapidly becoming available. Why then, are the majority of data collected by researchers still inaccessible? Impediments to data submission stem, in part, from the lack of suitable mechanisms to make it easy for an individual to submit datasets and metadata to a data centre or repository, and the lack of knowledge about the existence of appropriate data centres. Moreover, large, multiple-investigator projects often have little or no data management support to facilitate the organization and aggregation of data sets, and the subsequent transfer of data from the individual laboratories into accessible databases. Those relatively tractable issues, however, are arguably less a hurdle to data accessibility than the fundamental lack of incentives for researchers to provide their data for general use in the research community.

3. CURRENT PRACTICE

3.1 ACCESS TO DATA

There are two infrastructures operating in the oceanographic research community through which data and other research outputs may be obtained. Since the 1970s data centres such as the IODE network of national oceanographic data centres and the ICSU World Data Centre network have been ingesting data (in the sense of primary measurements of natural phenomena), adding value through standardisation, quality control and metadata enhancement. Their primary objectives have been to ensure long-term stewardship and that data may be reused with confidence decades after their collection – essential for detailed studies of global change. The problem is that only a small portion of oceanographic data collected make it into the centres and that those data that do get into data centres are regarded by parts of the scientific community as difficult to get out. The access difficulties are being addressed through the development of data delivery technology, such as SeaDataNet, but an answer to the issue of limited submission remains to be found.

In the 1990s a complementary approach, open access to research output, spawned a number of thematic repositories for research publications. Some like Los Alamos HEP EPrints (now ArKiv) and RePEc (Economics) have continued successfully, but in the main, subject-based repositories have been overtaken since 2000 by the implementation of Institutional Repositories (IRs). Under the Open Access Movement, these were set up to provide free access to research publications (without the need for subscriptions), but also to showcase each organization's research profile. Using custom-made Repository software and the Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH), publication citations and full text are made easily available for searching by Web search engines. However, organizations wish to capture and expose all their research output and so the IRs are being asked to extend their content by accepting multimedia and even datasets. In the UK a number of projects are looking at how IRs treat data (e.g. GRADE – geospatial data; CLADDIER – linking data and publications).

Calls for open access to data have been growing since the first Organisation for Economic Co-operation and Development (OECD) Official Statement in 1994 (see Box 1 for a list of statements in support of public access to data). There are now a number of position statements and policies, particularly from grant-awarding agencies, requiring grantees to deposit their datasets in publicly

accessible repositories. Apart from a few exceptions, deposit of datasets is not officially monitored, rewarded (or punished if not deposited), and therefore often not carried out by researchers.

Open-access repositories have the advantage that both getting data in and getting data out are both very straightforward and require very little human effort. Their limitation is that what comes out is exactly what goes in. Whilst this is not an issue for objects such as digital publications it can be a problem for data because unless the metadata standards are high it will be impossible in the future to understand what the measurements were, let alone make use of them. Years of experience receiving data submissions indicate that adequate metadata is very much the exception and not the rule.

Box 1 - STATEMENTS IN SUPPORT OF PUBLIC ACCESS (selection)

- Organization for Economic Co-operation and Development (OECD) Principles and Guidelines for Access to Research Data from Public Funding. 2007.
  Available: http://www.oecd.org/dataoecd/9/61/38500813.pdf
- European Union/European Science Foundation: European Research Council (ERC) Scientific Council Guidelines for Open Access, December 2007.
  Available: http://erc.europa.eu/pdf/ScC_Guidelines_Open_Access_revised_Dec07_FINAL.pdf
- European Research Advisory Board 2006
  http://ec.europa.eu/research/eurab/pdf/eurab_scipub_report_recomm_dec06_en.pdf
- European Geosciences Union via Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities 2003. Table of signatories @ 2008
  http://oa.mpg.de/openaccess-berlin/signatories.html
- National Science Foundation/National Science Board. Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century. NSB-05-40.
  Available: http://www.nsf.gov/pubs/2005/nsb0540/
- National Science Foundation Division of Ocean Sciences Sample and Data Policy 2003. NSF 04-004.
  Available: http://www.nsf.gov/pubs/2004/nsf04004/start.htm
- Global Biodiversity Information Facility www.gbif.org
- Ocean Biodiversity Informatics conference 2004 (Costello and Vanden Berghe 2006, Vanden Berghe et al. 2007)[1,2]
- National Institutes of Health (USA) Data Sharing Policy. 2006.
  Available: http://grants.nih.gov/grants/policy/data_sharing/
  http://grants.nih.gov/grants/policy/data_sharing/data_sharing_chart.doc
- Wellcome Trust Policy on Data Management and Sharing. 2007.
  Available: http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTX035043.htm
- ICSU Priority Area Assessment on Data and Information. 2004. Available:
  http://www.icsu.org/Gestion/img/ICSU_DOC_DOWNLOAD/551_DD_FILE_PAA_Data_and_Information.pdf
- UK Research Councils :
  BBRSC: http://www.bbsrc.ac.uk/publications/policy/data_sharing_policy.pdf
  ESRC: http://www.esrcsocietytoday.ac.uk/ESRCInfoCentre/Support/access/
  MRC : http://www.mrc.ac.uk/PolicyGuidance/EthicsAndGovernance/DataSharing/index.htm
  NERC: http://www.nerc.ac.uk/research/sites/data/policy.asp
- ICSU/CODATA. Available: http://www.codata.org/data_access/principles.html
- IOC Oceanographic Data Exchange Policy. 2007. Available: http://www.iode.org/policy
- Worldwide Funding Agencies via: Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities 2003. Table of signatories @ 2008
  http://oa.mpg.de/openaccess-berlin/signatories.html
- U.S. Global Change Research Program, Policy Statements on Data Management for Global Change Research.
  Available: http://www.gcrio.org/USGCRP/DataPolicy.html

3.2 Journal Publishers' practice

Publishers Policies on Data

In general, many publishers do not have a rigorous process to handle either submissions of data sets supporting the publication, or the reference to those data sets held in external data centres. However, a common principle in science, as stated by the journals *Science* and *Nature*, for example, is that scientists must make their data available for independent use, without restrictions, once it has been used in a publication. Publishers with such protocols have a variety of approaches to data hosting and data citation:

- **American Association for the Advancement of Science (AAAS)**: General Information to Authors. Data Deposition Policy in *Science*.
  Available: http://www.sciencemag.org/about/authors/prep/gen_info.dtl#dataavail
- **Nature Publishing Group**: Editorial policies, Availability of Data. Available:
  http://www.nature.com/authors/editorial_policies/availability.html
- **Ecological Society of America, Ecological Archives** -- Data Papers.
  Available: http://www.esapubs.org/archive/archive_D.htm
- **American Geophysical Union (AGU)**: Policy on Referencing Data in and Archiving Data for AGU Publications. Available: http://www.agu.org/pubs/inf4aus.html
- **Geological Society of America (GSA)** Position Statement. Open Data Access.
  http://www.geosociety.org/positions/pos7_data.pdf
- **Elsevier: Marine Micropaleontology,** Background Data Sets List.
  Available:
  http://129.35.76.177/wps/find/journaldescription.editors/503351/description#descriptionEarth
- **Earth System Science Data:** http://www.earth-system-science-data.net/general_information/index.html

3.3 Project data set publication on CD-ROM

During the 1990s it was common practice to publish the data sets from national and international projects on either CD-ROM or DVD. Examples include the UK North Sea Project[3], the European Ocean Margin Exchange Project[4,5], the BioMar project (Picton and Costello 1998)[6], the international World Ocean Circulation Experiment[7] and Joint Global Flux Study[8] projects. These products provided the projects with concrete deliverables emphasised by the strong project branding on disks and packaging.

The CD-ROMs and DVDs either produced or influenced by BODC and other data centres were typically designed with an emphasis on harmonisation and integration, and therefore did little to enhance the data originators' *Curriculum Vitae* through appropriate acknowledgement via data citation. Recommendations were given as to how usage of data from a specific originator should be acknowledged, but with some notable exceptions (JGOFS – USA, Germany, Australia and others), the recommended citation syntax did not map to a data subset that could be sensibly extracted. This prevented the construction of meaningful citations that could be used to enhance originators' CVs, as

such citations need to refer to something concrete. Consequently, whilst the publications delivered significant rewards to the projects they did little to enhance the CVs of the individuals who collected the data. For example, the OMEX I CD-ROM provided the guidance:

> *"Sufficient information has been provided in the Roll of Honour, in the data documentation and as originator codes tagged to the data for the originators to be identified. It is suggested that data be acknowledged by reference to the originator (e.g. Chou, 1997) with the CD-ROM cited as 'OMEX I Data Set, CD-ROM electronic publication, British Oceanographic Data Centre, Birkenhead, 1997."*

Other examples, such as the BioMar project CD-ROM, were designed as digital books, with an ISBN, citations for the complete product, and the five distinctive datasets within it cited as chapters (e.g. Picton et al. 1998, Connor et al. 1998, Kelly et al. 1998).[9,10,11]

4. RECOMMENDED SOLUTIONS

4.1 The Challenge

In addition to continuing the facilitation of data transfer, storage and preservation, researchers must be motivated to deposit their data within repositories where the data sets can be indexed and retrieved for use by the data originator(s) and by other researchers in either an Open Access or Controlled Access model.

Steps toward this objective include:

- Implement clear citation formats designed to provide
    - linkability to the data sets
    - clear acknowledgement to the originator(s).
- Ensure that published research derived from all or part of deposited data sets includes reference to those data sets using a standard citation format.
- Ensure that data sets refer back to the published articles (describing meaning and context).
- Educate both tenured and newly qualified researchers, and editors, as to the existence of repositories and to their responsibilities of ensuring data derived from public funding, or that support published papers, are generally accessible.
- Ensure that data held within existing repositories are citable, which a key missing component in most existing data systems.
- Illustrate the benefits available from data citations, as currently provided by article citations. Scientists are generally evaluated on the basis of their productivity in generating new knowledge as exemplified by their peer-reviewed publications and the perceived impact of those publications. This valuation can be extended to the "publication" of data sets within data repositories, and the subsequent citation of these data sets within the research literature. Data citations – in conjunction

with the traditional article citations – would enhance a researcher's record of achievement and provide a significant incentive to making their data sets available.

- Ensure there are adequate metadata associated with each data set so that the data are made visible to Web-based search engines. If researchers cannot easily locate the data most relevant to them, depositing data sets in huge data repositories would have minimal value.

- Ensure financial models/funding are in place for medium- and long-term preservation.

4.2 Citation of Data in Existing Repositories

The model of data centres being only a repository or an archive needs to change to being a data publisher; that is, they make data publicly available. Many data centres already do this (e.g. PANGAEA®), as do other data publishers such as GenBank, the Ocean Biogeographic Information System (OBIS), and Global Biodiversity Information Facility (GBIF). Although it has not been past practice, clearly data centres could provide data citations for new, and probably much of their currently held, data. The metadata describing datasets published by PANGAEA® and OBIS include citations in the normal fashion (i.e. author or editor, year, title, source, version, date accessed).

The fundamental data publication difficulty faced by many data centres, particularly IODE centres that developed during the 1980s, is that their ingestion model is to disaggregate datasets and then harmonise the component data into a common internal schema. Whilst this is well suited to serving synthesised data sets in response to ad-hoc queries based on user-supplied spatio-temporal co-ordinates, there is no strong dataset concept to provide 'hooks' for citations.

If they are to develop into data publishers these centres need to work in collaboration with their data suppliers to aggregate their data stock into citable datasets. Done properly, with sufficient attention to issues such as granularity, composition and data quality, this could develop existing repositories into effective digital libraries without compromising their pre-existing data delivery functionality.

A mechanism for data citation is required. The only mandatory constraint is that the citation contains all the information required to populate mandatory bibliographic metadata fields, but obviously maximum conformance to standards developed by the library community should be an objective. Whilst plaintext citations allow linkages to be forged between published text and datasets, they do not provide a guaranteed mechanism for dataset delivery. Universal Resource Locators (URLs) provide this in the short term but are fragile because they incorporate server naming and even file structures that inevitably change, leading to broken links. Instead, what is needed is a permanent label that is guaranteed to be resolvable into that latest URL for a dataset through a guaranteed on-line service, for example Digital Object Identifiers (DOIs).

4.3 Work flow for peer-reviewed data publications

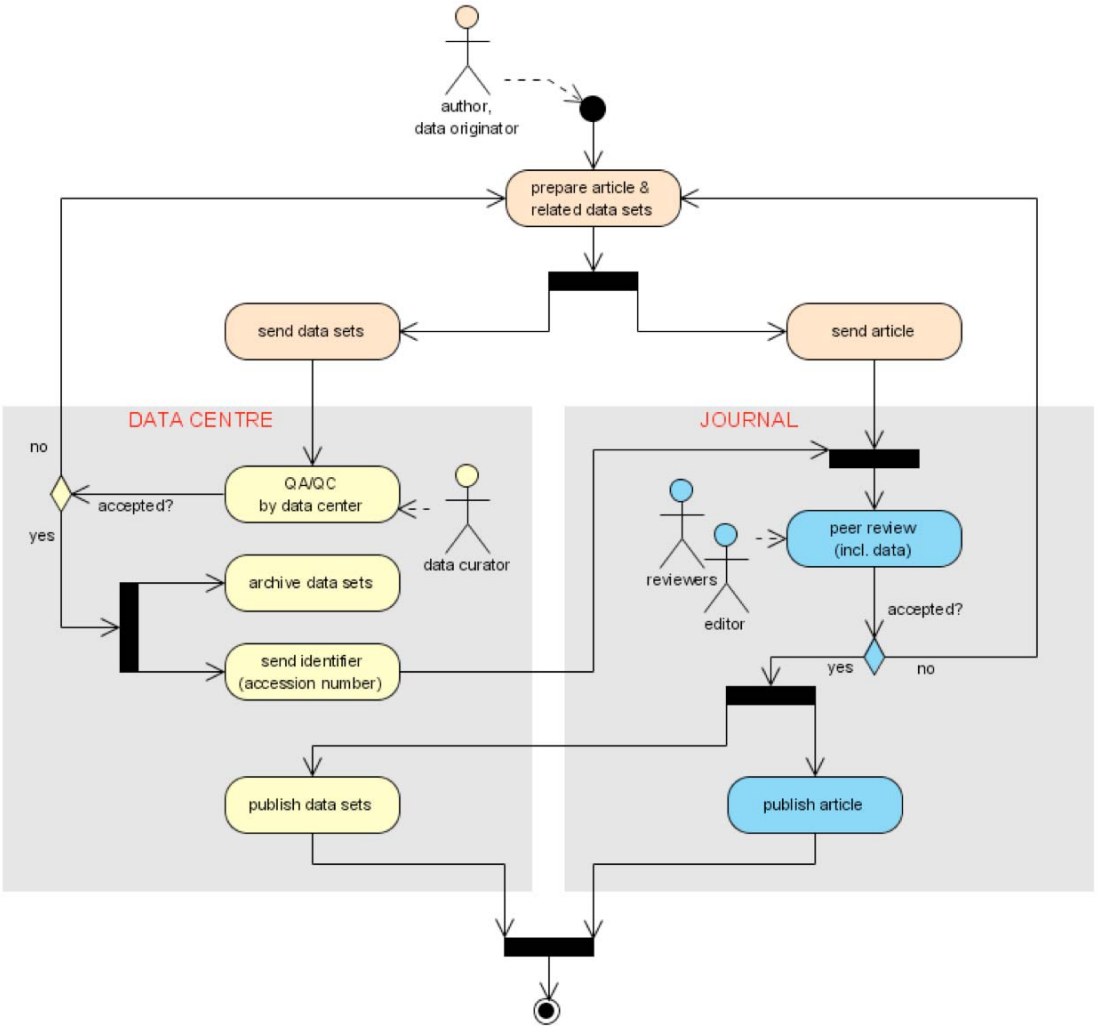

Figure 1: work flow diagram

The main objective of researchers is to communicate novel research through publications. For data publications two modes of operation have been discussed. Data can be published either as stand-alone publications or they can be published as supplements to traditional publications. The workflow, illustrated in the above figure, for both cases may be summarised as follows:

1. The author (data originator) sends the data (including the necessary metadata) to a certified data centre which has a clear mandate for a long-term data repository or data library and which is accepted by the journal publisher.
2. The data centre supplies a basic level of QA/QC. Data sets will be checked for adequate granularity, and completeness and correctness of metadata. Abnormalities (e.g. outliers) in the data will be identified and negotiated with the data originator.
3. If the data are accepted a persistent identifier (accession number) will be assigned and sent back to the data originator.
4. The article and the persistent identifier for the data are sent to the journal.

5. Peer review of the article will include review of related data sets. Reviewers will assess the validity of methods used, as well as the originality, lineage, and usability of the data. Usability covers issues such as format readability and the quality and completeness of metadata, particularly metadata describing what the measurements mean.
6. After acceptance, both article and data sets will be published and cross-referenced. The data sets can be referenced from the article either as a whole or eventually on finer levels (figures, tables)

All or some of the data sets stored in a permanent storage facility may be analyzed in support of a new research article. Resulting publications may be authored by the data contributor or anyone else; acknowledgement of the data originator as a citeable reference would be expected. Authorship of a data publication should involve data originators, except when not possible, for example, for legacy data.

Citations to supporting data external to the article are currently in non-standard formats, typically being a URL link. Optimally, data citations would be standardised to enhance discovery and referencing of data.

4.4 Providing a digital backbone for data journals

The publication process of datasets as 'data brief'-type papers involves elements of review such as standards conformance, data quality and metadata quality that form a part of data centre ingestion procedures. Consequently, there are benefits in developing collaborations between data journal editors and data centres. One mechanism for achieving this would be to use open archive e-Repository technology to establish input 'front-ends' to data centres that are discoverable through established standard metadata postings and capable of serving lodged datasets 'as is'. Dataset ingestion would then proceed as part of a review and dataset enhancement process involving the data centre, the data originator and the journal editor. Datasets posted would be assigned persistent identifiers that would tag all data entities in the data centres' systems related to the dataset.

If such infrastructure were established it could be harnessed to retrospectively publish, in close collaboration with data originators, data already held within the data centres that had not been used as the basis of scientific publications, such as monitoring datasets or delayed mode data streams from operational systems. This would go some way toward addressing the problem of inadequate reward for observational scientists who do make their data available through submission to data centres and provide incentive for their continuing this practice.

Infrastructure establishment obviously has cost implications. The model presented would require establishment of secure (i.e. backed up) Web-accessible storage plus adequate manpower within the data centres to handle ingestion. The latter is analogous to writing a blank cheque as there is no indication of the amount of new business that would result. Data centres would need to establish funding streams and negotiate service level agreements with the journals for this to become a practical proposition. Following the open-access model, data originators should pay for the publication. Charges, however, should be restricted to page charges for the article. At present, there is no foundation to pay for storage capacity used.

4.5 Providing a digital backbone for data sets in traditional papers

A clear need was identified to support accountability in the publication process by providing a means of storing the data used in the preparation of tables and figures in papers. This is accompanied by a

need to provide discoverable storage for 'supplementary data' that are increasingly accompanying digital publications whilst not appearing in print.

E-Repository technology again provides an obvious technological solution providing storage that may be discovered, served and cited in print. However, the case for coupling between the publication processes and data centres is less clear, bringing to the fore a role for a distributed network of institutional and thematic repositories. Resourcing is again an issue but in this case could be addressed through a storage charge model.

5. CONCLUSIONS AND WAY FORWARD

The workshop prepared a detailed work plan and associated deadlines as follows. In a few cases the meeting already started the work, but the following items map out the planned continued activity of SCOR and IODE on this topic.

1. **Prepare detailed workflow based upon preliminary draft and input from PANGAEA**

   A document will be provided that details the workflow for the data brief publication process. This document will be based upon the work flow of the existing PANGAEA repository. The document draft will be prepared by M. Diepenbroek and developed by the group (see Figure 1).

2. **Prepare draft terms of reference/guidelines for content contributors, repository input administrators, NODCs (and take into consideration certification issue)**

   (i) Terms of reference for National Oceanographic Data Centres related to data repositories for the data brief publication process of data:

   - Monitor submissions to designated e-repositories, most likely a part of the data centre, to identify targets for ingestion
   - Ingest identified target data
   - Collaborate with 'data journal' editors to provide technical review of published data sets as part of the centre's ingestion and quality-control procedures
   - Work with data originators, editors and reviewers to enhance 'data journal'-published dataset backbone objects, updating e-repository content. Enhancements would include metadata augmentation, quality control such as flagging, and standardisation such as translation into a standard format.

   (ii) Terms of reference for content contributors

   These will be prepared by Peter Wiebe. and Cisco Werner.

   (iii) Terms of reference for e-repository administrators operating as the digital backbone preserving data created during the scientific publication process.

   These will be prepared by Roy Lowry with collaboration from other group members.

3. **Draft deposit guidelines for content contributors**

   Increasingly, journals in all areas of science are requiring that all data that are required *"…to understand, assess, and extend the conclusions of the manuscript…"* (*Science*) be made available to any reader of the paper. But journals that have such mandates typically

offer limited or generalised guidance on what the nature of the data should be, what suitable repositories exist, or what the process is to accomplish a data submission. The following is an outline of the set of steps that can be carried out to fulfil the mandate.

1) Background data that were used to create summary tables, statistics and figures in the paper should be provided in suitable format to allow a reader to reproduce the summary table, statistics or figure.
2) Where substantial processing of an original data set is done to produce a figure or table, the modified data set should be provided. For example, for irregular spaced data, common techniques are to "krig" or objectively map the data to a regular spaced grid and then present the gridded data as a contour plot. The gridded data, a table of the parameter values used to do the objective mapping, an accounting of the software used, and adequate (good enough to allow reproduction of the processing if required) references to the original data should be provided.
3) In order to make the data discoverable and therefore useful for other purposes, additional information or metadata in standardized form may be needed.
4) Data should be submitted to a repository conforming to the IODE technical specification, and to the journal if editorial policies so require. A list of possible repositories will be developed.

A more detailed document will be drafted by Cisco and Joep and developed by the group.

4. Implement Pilot Project with established repositories

The meeting decided that a pilot project should be implemented. This pilot project will test the workflow of data through the journal publication process and into an existing repository. The pilot project will be coordinated by F. Werner and P. Wiebe. The work should be completed in time to report findings to the AGU Fall Meeting (15-19 December 2008). The work will also include user feedback and recommendations for required changes to the process (action item 1), draft terms of reference (action item 2), and the draft repository guidelines (action item 3).

5. **Develop technical description and requirements of repository system (including certification, citation formats, preferred file formats, metadata structures)**

- **Certification.** The meeting did not determine how certification would work, but proposed data repositories should provide the following information in writing:
  - Assurance of the permanence of their data holding, including how long the repository or associated data centre has been in existence, plans for archiving data holdings (including if data repository is closed some day), and plans to assign persistent identifiers.
  - Expected turn-around of accession numbers.
  - Services that will be provided.

- **Citation Formats**. These need to be developed based on existing work such as CLADDIER and the DOI infrastructure, and documented.

- **File Formats**. The constraints on what constitute acceptable formats for repository submission need to be discussed, agreed and documented.

- **Metadata**. The scope, content and format of mandatory discovery and usage metadata that should accompany repository datasets need to be discussed, agreed and documented.

A more detailed document will be drafted by **M. Diepenbroek, R. Lowry and P. Wiebe** and developed by the group.

6. **Convene meeting with publishers and editors**

Editors of ocean science journals and the publishers of these journals would have to be major participants in implementing the proposed system (or redesigning it in a way that would be more amenable to their capabilities and interests). Therefore, a meeting will be convened of the interested editors and publishers of major ocean science journals, both those published by societies and those published by commercial presses. An ideal opportunity for this meeting would be the Fall 2008 American Geophysical Union meeting in San Francisco, USA, and may be repeated later at a similar meeting in Europe, to involve editors and publishers representing biological journals, and those who are not planning to attend the AGU meeting. These meetings will expose the draft work flow process and attempt to develop agreement on how journals could better implement both data publications and access to data that underlie the figures, tables, and statistics in traditional journal articles. Ideally, a model "advice to authors" for data publications and supplemental data could be developed, which could be modified by individual journals as they desired. Some potential points to discuss include

- All journals should encourage submission to recognized publicly available open-access databases, of data supporting figures, tables, and statistics in their articles. There may need to be some negotiation with equipment providers for instrument outputs that are analyzed with proprietary software (e.g., software that converts acoustic returns to useable data). The idea in this point is that enough data be submitted that the reader can understand how primary data were converted to data actually used in the journal article, enhancing the ability of other investigators to understand (and repeat, if desired) the observations and experiments that are the focus of the paper.
- Agree to what would constitute an acceptable data repository and create a preliminary list. Draft criteria for acceptable repositories are given above; these should be discussed, modified, and approved by journals. Data repositories could be hosted by institutional e-repositories, national oceanographic data centres, and world data centres, but every centre would be required to meet the minimum standards.
- Journals should encourage data publications in their subject area and publishers should consider establishing new data journals. Current practice varies, as described earlier. There is one new data journal for Earth system science, *Earth System Science Data*, but this journal is still developing its processes. Whether it is desirable to establish data journals when existing data centres may better perform this role and facilitate data integration across data sets, is questionable.
- These data publications and supplemental data reports should be in machine-readable formats, following a small number of standard formats. Data centres and contributing scientists should advise on this. Data repositories, as defined herein, will be archives that do not perform a significant amount of value-added processing, as a data centre would, so it will be important that a small number of acceptable formats for data submission be agreed. This will make it possible to develop software that any user could have on their computer to access the data in repositories. It is especially important that the data be available in machine-readable format and not, for example, PDF format.

- Agreement on pointers from papers to data sources, and data sources to papers and citation formats for data in journal articles.
- Discuss peer review of supplemental data and data publications. Peer review is an important part of the work flow for data publications and a standard process for such peer review could help the journals create instructions for reviewers. The data underlying traditional articles is not normally peer reviewed, but reviewers could examine the data if they desired.

These will be augmented by other points raised by the editors and publishers before and at the meeting. Practical arrangements for the AGU meeting (**15-19 December 2008)** will be coordinated by **F. Werner and E. Urban.**

7. Follow-up meeting to discuss implementation

A follow-up meeting is tentatively planned for **early 2009** to evaluate dissemination activities and the results of the meeting(s) with editors and publishers. Date and venue will be decided later, depending on availability of the experts and financial resources. The practical details for the meeting will be handled by **E. Urban and P. Pissierssens.**

8. Dissemination and promotion

The ideas proposed in this report are not entirely new, but would involve a significant change in the culture of the ocean sciences community. A number of approaches will be used to promote the suggested changes. First, we will seek to disseminate information through a variety of channels through which ocean scientists normally receive information. These will include a short article in *EOS*; a presentation at AGU, European Geophysical Union, American Society for Limnology and Oceanograhy, and potentially other major meetings; information presented on the SCOR and IODE Web sites and newsletters; and focused meetings with groups that need to be involved in implementation. Special attention should be given to training graduate students about these concepts. This can be accomplished through a brochure, email lists, scientific meetings and conferences, summer schools, publishers, and other organizations and their related activities (e.g., the new Nippon Foundation (NF)/Partnership for Observation of the Global Oceans (POGO) Centre of Excellence in Observational Oceanography at the Bermuda Institute of Ocean Sciences). The IOC/IODE OceanTeacher system could also be an appropriate tool to disseminate training materials related to the repository. Several SCOR projects have data management committees and this report will be transmitted to them. One project, the Integrated Marine Biogeochemistry and Ecosystem Research (IMBER) project will hold data management training in conjunction with its November 2008 conference (see https://www.confmanager.com/main.cfm?cid=1185&nid=9105) and we will encourage them to discuss the data publication issue in that special session. Another aspect of dissemination will be to transmit the information beyond ocean science. As part of the new project “Pan-European Species Infrastructure” (PESI), M. Costello will be engaging with journal publishers to promote the use of biodiversity data standards for data appendices, the use of online data systems such as OBIS and GBIF, and the citation of datasets.

It was further agreed that a short report on the SCOR/IODE meeting will be prepared (500 words) for EOS by Ed Urban, Roy Lowry and Peter Pissierssens.

6. References

1. Costello, M.J., Vanden Berghe E. 2006. "Ocean Biodiversity Informatics"enabling a new era in marine biology research and management. MarineEcology Progress Series 316, 203-214. http://www.int-res.com/abstracts/meps/v316/.
2. Van den Berghe, E., W. Appeltans, M.J. Costello, P. Pissierssens (Eds). (2007) Proceedings of 'Ocean Biodiversity Informatics': an international conference on marine biodiversity data management Hamburg, Germany, 29 November - 1 December, 2004. Paris, UNESCO/IOC, VLIZ, BSH, 2007. vi + 192 pp. (IOC Workshop Report, 202) (VLIZ Special Publication, 37)
3. North Sea Project Data Set (1992). British Oceanographic Data Centre, Birkenhead, United Kingdom. CD-ROM.
4. OMEX I Data Set (1997). British Oceanographic Data Centre, Birkenhead, United Kingdom. CD-ROM.
5. OMEX II Data Set (2002). British Oceanographic Data Centre, Birkenhead, United Kingdom. CD-ROM.
6. Picton, B.E. and Costello M. J. (editors) 1998. *The BioMar biotope viewer: a guide to marine habitats, fauna and flora in Britain and Ireland*, Environmental Sciences Unit, Trinity College, Dublin. ISBN 0 9526 735 4 1
7. WOCE global data, Edition 3.0. Southampton, UK, World Ocean Circulation Experiment (WOCE). DVD x 2. (WOCE Report 180/02).
8. International Collection of JGOFS (Joint Global Ocean Flux Study) Volume 2: Integrated Data Sets (1989-2003). Sieger, R et al., JGOFS DMTT & IPO, WDC-MARE Reports 0003, 15 pp with CD-ROM, World Data Center for Marine Environmental Sciences, Bremen/Bremerhaven, Germany, 2005.
9. Picton B.E., Emblow, C.S., Morrow, C.C., Sides, E.M., Tierney, P., McGrath, D., McGeough, G., McCrea, M., Dinneen, P., Falvey, J., Dempsey, S., Dowse, J. and Costello, M. J. 1998. Marine sites, habitats and species data collected during the BioMar survey of Ireland. In: Picton, B.E. and Costello M. J. (eds), *The BioMar biotope viewer: a guide to marine habitats, fauna and flora in Britain and Ireland*, Environmental Sciences Unit, Trinity College, Dublin.
10. Connor, D.W., Brazier, D.P., Dalkin, M.J., Hill, T.O., Holt, R.H.F., Northen, K.O. and Sanderson, W.G. 1998. Marine Nature Conservation Review: marine biotope classification for Britain and Ireland, Version 97.06. In: Picton, B.E. and Costello M. J. (eds), *The BioMar biotope viewer: a guide to marine habitats, fauna and flora in Britain and Ireland*, Environmental Sciences Unit, Trinity College, Dublin.
11. Kelly, K. S., Costello, M. J., Baxter, P. W. and Picton, B. E. 1998. A bibliography of Irish marine literature from 1839-1997. In: Picton, B.E. and Costello M. J. (eds), *The BioMar biotope viewer: a guide to marine habitats, fauna and flora in Britain and Ireland*, Environmental Sciences Unit, Trinity College, Dublin.

ANNEX I

AGENDA OF THE MEETING

1. OPENING OF THE MEETING

2. CURRENT STATUS OF DATA CITATION/DATA PUBLICATION IN OCEANOGRAPHY

2.1 DOIs at WDC-MARE (Michael Diepenbroek, WDC-MARE)
2.2 DOIs at U.S. Department of Energy (Christopher Lenhardt, Oak Ridge National Laboratory, Distributed Active Archive Center)
2.3 The GenBank experience (Peter Wiebe, WHOI)
2.4 Journal-based data reports (Cisco Werner, Rutgers University)
2.5 Data citation and the publishers (Craig Emerson, ProQuest)
2.6 Linking data and publications: The CLADDIER Project experience (Pauline Simpson)
Motivating data publication (Mark Costello, University of Auckland)
2.7 Discussion

3. ANALYTIC SUMMARY

4. DISCUSSIONS

5. CONCLUSIONS AND WAY FORWARD

ANNEX II

LIST OF PARTICIPANTS

EXPERTS

Dr Mark COSTELLO
Associate Professor
Leigh Marine Laboratory,
University of Auckland
P.O. Box 349
Warkworth, New Zealand
Tel (64)(9)3737 599 x 83608
Fax (64)(9)422 6111
e-mail: m.costello@auckland.ac.nz

Dr Michael DIEPENBROEK
Managing Director
WDC/MARE
Leobener Strasse
Bremen, Germany
Tel: +49 421 218 65590
e-mail: mdiepenbroek@pangaea.de

Dr Craig EMERSON
Vice President, Editorial Operations
ProQuest
P.O. Box 1346
Ann Arbor. MI 48106-1346
United States
Fax: 1 3019616744
e-mail: Craig.Emerson@ProQuest.com

Mr Jan HASPESLAGH
Librarian
Flanders Marine Institute/ Vlaams Instituut voor de Zee (VLIZ)
Vismijn Pakhuizen 45-52
Oostende 8400
Belgium
Tel: +32-59-342130
Fax: +32-59-342131
e-mail: janh@vliz.be

Mr Francisco HERNANDEZ
IT Manager, Datacenter
Flanders Marine Institute/ Vlaams Instituut voor de Zee (VLIZ)
Vismijn Pakhuizen 45-52
Oostende 8400
Belgium
Tel: +32-59-342130
Fax: +32-59-342130
e-mail: francher@vliz.be

Mr W. Christopher LENHARDT
Informatics Scientist
Environmental Sciences Division
Oak Ridge National Laboratory (ORNL)
P.O. Box 2008
Oak Ridge TN 37831-6036
United States
Tel: 1 865 574 6332
Fax: 1 865 241 3685
e-mail: lenhardtc@ornl.gov

Dr Roy LOWRY
British Oceanographic Data Centre (BODC)
Joseph Proudman Building
6 Brownlow Street
Liverpool L3 5DA
United Kingdom
e-mail: rkl@bodc.ac.uk

Mr Richard Pepe
Fishery Information Officer (Editor-in-Chief ASFA)
Food and Agriculture Organization of the United Nations (FAO)
Fisheries and Aquaculture Department
Fisheries and Aquaculture Economics and Policy Division (FIE)
Fisheries and Aquaculture Information and Statistics Service (FIES)
Via delle Terme di Caracalla
00153, Rome
Italy
Tel: (39)[06]57056380
Fax: (39)[06]57053605

e-mail: richard.pepe@fao.org

Ms Pauline SIMPSON
Program Coordinator
Central Caribbean Marine Institute (CCMI)
P.O. Box 10152
Grand Cayman KY1-1002
Cayman Islands
Tel: +[44] 7715 113930
e-mail: psimpson07@aol.com

Mr Joep VERHEGGEN
Director of Product Management
Academic and Government, Elsevier
Elsevier B.V.
Radarweg 29
1043 NX Amsterdam
The Netherlands
Tel: 31 20 4852501
Fax: 31 20 4583354
email: j.verheggen@elsevier.com

Prof Francisco WERNER
Director
Institute of Marine and Coastal Sciences
71 Dudley Road
Rutgers University
New Brunswick, NJ 08901
United States
Phone: +1-732 932-6555, Ext. 509
Fax: +1-732 932-8578
e-mail: cisco@marine.rutgers.edu

Dr Peter WIEBE
Scientist Emeritus
Biology, MS#33
Wood Hole Oceanographic Institution
Woods Hole, MA 02543
United States
Tel: 1-508-289-2313
Fax: 1-508-457-2169
e-mail: pwiebe@whoi.edu

IOC SECRETARIAT

Mr Peter PISSIERSSENS
IODE Programme Coordinator
Head, IOC Project Office for IODE
Wandelaarkaai 7
8400 Oostende
BELGIUM
Tel: +32-59-34 01 58
Fax: + 32-59-34 01 52
e-mail: p.pissierssens@unesco.org

SCOR SECRETARIAT

Dr Ed URBAN
Executive Director
Scientific Committee on Oceanic Research
Robinson Hall
College of Marine and Earth Studies
University of Delaware
Newark, DE 19716
United States
Tel: 1-302-831-7011
Fax: 1-302-831-7012
e-mail: ed.urban@scor-int.org

ANNEX III

TERMINOLOGY DEFINITIONS

Repository: a repository is a central place in which an aggregation of data is kept and maintained in an organized way, usually in computer storage. Depending on how the term is used, a repository may be directly accessible to users or may be a place from which specific databases, files, or documents are obtained for further relocation or distribution in a network. A repository may be just the aggregation of data itself into some accessible storage location, or it may also imply some ability to selectively extract data. (from: whatis.com)

e-repository: A computer storage resource for digital objects (files of collections of files) that are addressed by dataset references, usually in the form of a citation.

e-repository administrators: the group of people responsible for the management of an e-repository.

Data: "research data" are defined as factual records (numerical scores, textual records, images and sounds) used as primary sources for scientific research, and which are commonly accepted in the scientific community as necessary to validate research findings. A research data set constitutes a systematic, partial representation of the subject being investigated.

Data publication (verb): process through which data are fixed and made citable and retrievable over the long term and may imply there has been a quality-control process. (noun) a peer-reviewed, citeable publication that focuses on data, with a short textual explanation of the data set, without interpretation.

Supplemental data: any digital object (data, multimedia, …) associated with a published journal article (but not embedded in the article).

Data centre: an organisation that collects research data and manages them with the objective of facilitating their re-use, such as providing baselines for studies of change, decades after their collection. Data centres are generally concerned with observational data, add value to data through quality assurance and metadata enhancement and have an operational model based on data harmonisation into a common schema.

National Oceanographic Data Centres: an international network of data centres developed by IOC to form IODE.

Content contributors: scientists who provide data to data centres or digital objects to e-repositories.

ANNEX IV

LIST OF ACRONYMS

AAAS	American Association for the Advancement of Science
AGU	American Geophysical Union
ASLO	American Society of Limnology and Oceanography
BODC	British Oceanographic Data Centre
BBSRC	Biotechnology and Biological Science Research Council
CLADDIER	Citation, Location, and Deposition in Discipline & Institutional Repositories
DOI	Digital Object Identifier
EGU	European Geosciences Union
ERC	European Research Council
ESRC	Economic and Social Research Council
GBIF	Global Biodiversity Information Facility
GRADE	Grading of Recommendations Assessment, Development and Evaluation
GSA	Geological Society of America
ICSU	International Council for Science
IOC	Intergovernmental Oceanographic Commission
IODE	International Oceanographic Data and Information Exchange
IMBER	Integrated Marine Biogeochemistry and Ecosystem Research
JGOFS	Joint Global Ocean Flux Study
IR	Institutional Repository
MRC	Medical Research Council
NERC	Natural Environment Research Council
NF	Nippon Foundation
NODC	National Oceanographic Data Centre
NSF	National Science Foundation
OAI-PMH	Open Archive Initiative Protocol for Metadata Harvesting
OBIS	Ocean Biogeographic Information System
PESI	Pan-European Species Infrastructure
OECD	Organisation for Economic Co-operation and Development
POGO	Partnership for Observation of the Global Oceans
QA	Quality Assurance
QC	Quality Control
SCOR	Scientific Committee on Oceanic Research
URL	Universal Resource Locator
WDC-MARE	World Data Centre for Marine Environmental Sciences

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