SCOR/IGBP Meeting on Data Management for International Marine Research Projects[1]

Introduction

The Scientific Committee on Oceanic Research (SCOR) and International Geosphere-Biosphere Programme (IGBP) convened a meeting on Data Management for International Marine Research Projects on 8-10 December 2003 in Liverpool, UK. Meeting participants included representatives (both data producers and data managers) from international projects and programmes (CLIVAR, GEOHAB, GLOBEC, IGBP, IMAGES, IMBER, IODE, JGOFS, LOICZ, OBIS, SCOR, SOLAS, WDCs and WOCE) and data managers from national data centres (BODC, Indian NODC), see appended list). Dr. Roy Lowry of the British Oceanographic Data Centre (BODC) convened the meeting at the Foresight Centre, University of Liverpool. SCOR and IGBP thank the U.S. National Science Foundation and BODC for their support of this meeting.

Three important products resulting from this meeting are presented in this document:

(1) a series of recommendations based on reports from marine research projects, and presentations and discussions at the meeting;
(2) agreement on, and modifications to, recommendations from a working group on Oceanographic Data Management held at the IGBP Congress in Banff in June 2003; and
(3) guidelines for development of project data policies.

These three products follow. The session summaries (Appendix I), presentation documents (Appendix II), meeting agenda (Appendix III), and participants' list (Appendix IV) are also available on the activity Web page (see www.jhu.edu/scor/DataMgmt.htm). This meeting was designed to fulfil one of the recommendations from the Banff meeting and to extend the work started in that session.

Recommendations

The following recommendations were distilled by the rapporteurs from the discussion sessions during the meeting.

Session on Preceding Work

- The report from the data management session at the IGBP Congress at Banff in June 2003 should be taken forward as a recommendation from the

[1] This document is based on work partially supported by the U.S. National Science Foundation to the Scientific Committee on Oceanic Research (SCOR) under Grant No. 0326301. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. National Science Foundation (NSF).

Liverpool meeting after an agreed set of modifications has been incorporated. The modified document is included below.

- An information resource to support the development of data management in new projects should be established and maintained. A Web site would seem the most appropriate vehicle for such a venture. SCOR has offered to host this facility, at www.jhu.edu/scor/DataMgmt.htm

Session on New and Developing Projects

- Established data management expertise and techniques in the marine science community do not address the management of data that are not geospatially referenced or socio-economic data. Projects need to address these data types through the composition of their Data Management Committees.
- Data Management Committees from different marine projects would benefit from joint meetings to ensure common solutions to common problems. A mechanism is required to facilitate such meetings.
- Data Management Committees should have three areas of responsibility:

  (1) ensure that data are available for project scientific purposes and that data management meets the present scientific need of the project without compromising future needs,
  (2) oversee the compilation of data from individual principal investigators (PIs) and national projects into a long-term, integrated data set that is submitted to an appropriate data archive and may be published in CD-ROM or DVD format, and
  (3) address the involvement in project data exchange activities of scientists without access to effective data management infrastructure.

Session on the WOCE experience

- The Data Assembly Centre (DAC)[2] model adopted by the World Ocean Circulation Experiment (WOCE) is applicable to predictable data management requirements such as CTD data management and assembly (but not quality control) of water bottle data sets. Such infrastructure could be shared among projects. However, the concept cannot be extended to cover the full range of parameters measured by biogeochemical and ecological projects.

Session on Metadata[3] Management

- Metadata management should be decoupled from data management, with the International Project Office (IPO) taking the lead role in metadata catalogue assembly.

[2] Data Assembly Centres assemble a restricted data stream (such as sea level data, current meter data or CTD data) from all sources of that data type within a project. DACs are generally established in centres with an established reputation for handling the type of data concerned and generally represent in-kind donations by national infrastructure to the international programme.

[3] Metadata are information about data, including information that allows data sets to be located (discovery metadata: what was measured, when and where), information that enhances human understanding of the data and the uses to which it can be put (semantic metadata) and information that allows software agents to access the data (technical metadata).

- Endorsement of scientific activity by a project requires metadata to be submitted on the shortest possible timescales. This requirement should be clearly stated in the project data policy, including mechanisms for metadata submission and sharing.
- Common metadata standards should be adopted across projects to facilitate sharing of metadata through catalogue interoperability. The Directory Interchange Format (DIF) developed by the Global Change Master Directory (GCMD) is a suitable standard for cataloguing datasets and has established storage and query infrastructures.
- Project metadata catalogues should be combined through distributed networking or even physically combined, when required by technical considerations.

Session on National Data Management Infrastructure

- The co-operative data management strategy operated by BODC project data management and IMAGES data management (sometimes termed "end-to-end" data management, which is the phrase used elsewhere in this document)[4] is a useful concept and should be reproduced and adapted. It may be implemented with either the data manager and database infrastructure in the same organization, such as BODC, or in different organizations, such as the IMAGES data manager and WDC-MARE. Combinations of these two modes of operation allow a totally scalable infrastructure to be developed. The data manager role also could be operated by a small- to medium-sized commercial enterprise (SME).
- Procedures could be developed to operate end-to-end data management in countries without an adequate data management infrastructure in collaboration with an established data centre.

Session on Data Policies

- As SCOR and IGBP are ICSU bodies, SCOR and IGBP projects must adopt the ICSU principle of free and open data exchange.
- There are many data management policies that could be used as the basis of an IGBP/SCOR template (see pages 7-9) and adapted to the specific needs of each project.
- Project SSCs should decide the rules for data access within and between projects.
- Data Management Committees should monitor adherence to their policies and report breaches to be dealt with on a case-by-case basis by the SSC.

Session on Technical Aspects of Data Management

- It is essential that projects identify and universally adopt appropriate data and metadata standards at the start of the project.

[4] Project data management or end-to-end data management is characterized by the involvement of the project data manager from the beginning of the project, in planning how the data will be collected, shared, and archived. The data manager may be involved in planning the research, participate in research cruises, help participating scientists format their data and train them in methods to use project data, and ensure that project data are archived in appropriate national and international data archives.

- A technical forum is required to ensure that the compatible standards are maintained across projects.
- Meeting participants expressed concern at the number of distributed data systems[5] currently being independently developed to very similar specifications. A meeting of distributed system developers is recommended to ensure interoperability[6] among these systems.
- The technology needs of developing countries may be more effectively addressed through infrastructure development rather than through restriction of technological developments elsewhere. There is a clear need for reliable high-bandwidth network capacity across the globe.
- There is a need to develop infrastructure to ensure the long-term availability of real-time data that are currently displayed on the Web for a limited time and then destroyed.

Session on Data Submission to World Data Centres (WDCs)

- A peer-reviewed dataset publication infrastructure should be established and efforts made to initiate culture change in marine sciences to raise the status of these publications as output performance indicators.
- Data quality control should be the result of a partnership, with data originators, data users, and data centres (national and world) each playing a role.
- Countries that do not have a national oceanographic data centre within the IODE network should establish a national data co-ordinator. Countries should inform their international projects of their appropriate national data centre or national data contact, and should secure the adequate involvement of their national data centre in national and international management of project data.
- If data coverage gaps within the WDC system are identified, then a dialogue among the projects, the relevant WDCs, and the Intergovernmental Oceanographic Data Exchange (IODE) is recommended.
- The acquisition of data by WDC should be a partnership between IODE network, the WDCs, and project data managers.

Session on Funding Data Management

- The proportion of the total project science budget (including platform costs) required for end-to-end project data management, project data services, and assuring the long-term stewardship of the project data is approximately 10%.
- Operating a project metadata catalogue should be considered a core activity of the IPO and requires a minimum of one-half of the time of a full-time employee.

[5] Distributed data systems are systems where data held in multiple databases in multiple locations are accessed through a common user interface, commonly termed a portal. Examples include OpenDAP (DODS), LAS, Mercury, and Thredds.

[6] Interoperability is the ability to seamlessly access metadata or data held in multiple databases, exactly as though they were from a single database.

Oceanographic Data Management: Recommendations from Working Group B2 at the Banff Congress (June 2003)

Affirmed and modified by participants of the SCOR/IGBP Meeting on Data Management for International Marine Research Projects, 8-10 December 2003

Session Summary

The session opened by declaring its primary objective to be to help the new projects, such as SOLAS, IMBER and the SCOR/IOC GEOHAB initiative, to develop their data management plans.

The data management of the mature programmes JGOFS, WOCE, LOICZ and GLOBEC was reviewed. It became clear from this that the following actions are extremely beneficial to projects:

- Establishment of a Data & Information Management Unit at the outset.
- Development of scalable data management
- Adoption of standards to facilitate interoperability of data and information, while allowing for evolution of techniques during the programme
- Utilisation of existing infrastructure but with additional resources to ensure it fulfils international rather than national specifications and standards
- Provision of services and data access that match the needs of scientists and other end users
- Provision of data through both a leading edge technology and a universally available technology
- Development of a close working relationship between data managers and scientists through means such as "end-to-end" project data management and the provision of data access tools

Some generic data management issues were then examined:

- The form and content of a "data policy."
- The role of developing technologies, such as the development of seamlessly integrated distributed databases
- Areas where oceanographic data managers need to look for new techniques, such as socio-economic data, bio-informatics and non-spatial data, for example, mesocosm and other experiments

Strategy scenarios to bridge the gap between data at the "PI" level and a complete, fully integrated and documented data set were then examined.

The session was concluded by drawing together the following recommendations:

Recommendations for New IGBP Oceanographic Programmes

1. Projects should establish a data policy at the outset to address the following issues:

    - Data sharing within the programme, between programmes and the entry of data into the public domain.
    - Data content and quality issues.
    - Long-term security of the data.

2. All new programmes should dedicate resources to the development of a project meta-database that will form the project data inventory. This should conform to appropriate international standards (e.g., ISO19115 for spatially referenced data) to facilitate integration and exchange of information between programmes. The IPO should ensure that a structure is created and implemented, appropriate to the needs of the project.

3. Projects should establish a data management working group such as the JGOFS Data Management Task Team or the WOCE Data Products Committee. Past experience has shown that these groups are more effective if they comprise data originators, data managers and data users.

4. National or project science programmes should address data management in a credible manner, including allocation of appropriate resources and giving consideration to capacity building, if appropriate.

5. Attention should be given to developing incentives for scientists to submit/share their data, for example, by offering tools such as modelling, plotting, and cartographic representation of data.

Recommendations for Further Work

1. A data policy template should be developed to assist programmes with the compliance with recommendation (1) above.
2. IGBP should work together with other international organizations to promote a culture where datasets are regarded as citable entities that are recognized as important scientific outputs.

Roy Lowry, British Oceanographic Data Centre (Chair)
Bernard Avril, JGOFS IPO (rapporteur)
23/06/2003
Revised 10/12/03

Data Policy Template for IGBP and SCOR Marine Projects

Scientific data and information derived from large-scale research projects with oceanic components are critical to project success and are an important legacy of these projects. Project data should be available for assessment and use by independent scientists, including, initially, other project scientists and later by external scientists. To ensure long-term survival, integrity, and availability of project data and models, a workable plan, policy, and associated infrastructure must be established early in the life of a project. Project data, as well as model code and model output, must be made available to the community.

A data management policy and plan should (1) encourage rapid dissemination of project results; (2) ensure long-term security of key project data, as well as model-related information; (3) protect the rights of the individual scientists; (4) treat all involved researchers equitably; and (5) reward openness. IGBP and SCOR affirm the data policy of their parent organization, the International Council for Science (ICSU):

> "ICSU recommends as a general policy the fundamental principle of full and open exchange of data and information for scientific and educational purposes." [ICSU General Assembly Resolution 1996]

Participants at the December 2003 meeting on Data Management for International Marine Research Projects recommend that all IGBP/SCOR large-scale marine research projects adopt the following essential elements in their data policies. Also listed are additional considerations for the development of project data management systems.

<u>Essential Data Policy Elements</u>

- Project endorsement requires a credible commitment to the timely submission of data to a project-approved database to ensure long-term archiving of the data.
- Discovery Metadata (what was collected where, when and by whom) should be submitted by project scientists to the International Project Office on the shortest feasible time scales. Failure to do should be considered reason to remove project endorsement.
- Model code and documentation, initialisation, boundary conditions, data used to force the model system, and output resulting in published results ("definitive runs") must be submitted to project-approved databases in forms which allow assessment of key findings.
- Timelines for data and model sharing, as well as protocols associated with intellectual property rights of different data types and models, should be defined. Currently accepted guidelines are that data should enter the public domain after a maximum of two years after data become available to the PI.
- Quality control of metadata, data and model output needs to be addressed.
- Each project should form and support a Data Management Committee. The three primary functions of Data Management Committees are to:

(1) ensure that data are available for project scientific purposes and that data management meets the present scientific need of the project without compromising future needs,
(2) oversee the compilation of data from individual principal investigators (PIs) and national projects into a long-term, integrated data set that is submitted to an appropriate data archive and may be published in CD-ROM or DVD format, and
(3) address the involvement in project data exchange activities of scientists without access to effective data management infrastructure.

- Projects must adopt or establish a credible data management infrastructure.
- Projects should adopt metadata standards (content and controlled vocabularies[7]) and agreed data formats both within and among projects to facilitate data interoperability.
- Project Data Management Committees should consider how to get appropriate project data into operational data streams[8] and appropriate operational data streams into the project domain.

Additional Considerations

Project SSCs and Data Management Committees should create their project data policy, considering the following issues.

The project SSC should:

- Create a Data Management Committee with adequate representation of project science, a balance between project scientists (including modellers), national and international project data managers, and consideration of outreach functions to countries without data centres.
- Consider providing access to project-related publications through a publication database, such as that used by GLOBEC.

The project Data Management Committee[9] should:

- Develop a process to ensure that metadata and data are submitted, monitor the compliance of project scientists to the policies, and refer failure in compliance to the project SSC.
- Specify how project data will be quality controlled.
- Specify incentives to encourage project scientists to submit metadata and data to the IPO and a long-term data repository, respectively. ("One carrot is worth ten sticks.") These incentives may include citation of data in a peer-reviewed journal, access to other project data during "an embargo period" before public access, tools for use of data in the data archive (e.g., data merging, plotting,

[7] Metadata vocabularies are controlled lists of words or phrases that are used to populate metadata fields in place of free text to ensure computer searches are not compromised by problems such as spelling variations.
[8] Operational data streams are data that are available on a regular basis from routine observing systems, such as Argo floats, sea level networks, and telemetered data buoys.
[9] Where modelling committees exist, these should be consulted in relation to model-specific aspects of data policy.

spatial visualisation and modelling tools), and help from international data managers in submitting data, accessing data, and using analysis tools. Proper incentives will reduce the efforts needed by data managers to get data into project data systems and increase participation in the project.

- determine the variables most likely to be measured and the expected data volumes, and specify project data products.
- address how non-geo-referenced, socioeconomic, and other non-conventional data will be handled.
- consider setting up a DAC, either project-specific or shared among projects, for data that can be handled in this way. The DAC may be set up along the lines of project data streams (e.g., CTD data, bottle data) and/or the more traditional single parameter DAC (i.e., the DACs used by WOCE and CLIVAR).
- consider whether to submit DIFs to GCMD as a means to provide access to project metadata.
- consider making species-specific data available through OBIS.
- create a mechanism to interact regularly with representatives of related project Data Management Committees to develop common approaches and procedures to share data.

Project SSCs and Data Management Committees should work together to

- specify how project models and data will be made available both to scientists with leading-edge technology and with unreliable access to even basic access methods. The project should also present plans for training developing country scientists in techniques for data access and use.
- develop plans to bring together data providers and data managers, considering how "project data management" principles could be applied to each project.

Acronyms

BODC	British Oceanographic Data Centre
CLIVAR	Climate Variability and Prediction project
DAC	Data Assembly Centre
DIFs	Directory Interchange Formats
GCMD	Global Change Master Directory
GEOHAB	Global Ecology and Oceanography of Harmful Algal Blooms programme
GLOBEC	Global Ocean Ecosystem Dynamics project
ICSU	International Council for Science
IGBP	International Geosphere-Biosphere Programme
IMAGES	International Marine Aspects of Global Change project
IMBER	Integrated Marine Biogeochemistry and Ecosystem Research project
IODC	Indian Oceanographic Data Centre
IODE	Intergovernmental Oceanographic Data and Information Exchange
IPO	International Project Office
JGOFS	Joint Global Ocean Flux Study
LOICZ	Land-Ocean Interactions in the Coastal Zone project
NODC	National Oceanographic Data Centre

OBIS	Ocean Biogeographic Information System
PI	principal investigator
SCOR	Scientific Committee on Oceanic Research
SSC	Scientific Steering Committee
SME	small to medium-sized commercial enterprise
SOLAS	Surface Ocean – Lower Atmosphere Study
WDC	World Data Centre
WDC-MARE	World Data Centre for Marine Environmental Sciences
WOCE	World Ocean Circulation Experiment

Meeting Participants

Name	Project/Organization
Dawn Ashby	GLOBEC IPO
Bernard Avril	JGOFS IPO
Geoff Boxshall	OBIS
Wendy Broadgate	IGBP Secretariat
Juan Brown	BODC
Howard Cattle	CLIVAR IPO
Wolfgang Fennel	GEOHAB
Julie Hall	IMBER
Katy Hill	CLIVAR IPO
Roy Lowry	BODC
Liana Talaue-McManus	LOICZ
Lesley Rickards	IODE
Stefan Rothe	IMAGES
Casey Ryan	SOLAS IPO
Jaswant Sarupria	IODC
James H. Swift	WOCE
Ed Urban	SCOR Secretariat
Douglas Wallace	SOLAS
Ferris Webster	ICSU WDCs