

## **Data Management in IMBER**

### **Report from the 1<sup>st</sup> IMBER Data Management Committee (DMC) Meeting**

10-11 June 2007, Victoria, Canada

#### **Contents**

1. Executive summary .....	p3
2. Preamble.....	p3
3. Objectives of the first DMC meeting.....	p3
4. IMBER Goals and themes.....	p4
5. IMBER data management - introduction and overview (ambitions).....	p4
5.1 Introduction	
5.2 IMBER data	
5.3 Oversight of IMBER Data Management	
6. End to end data management - role of Data Specialists.....	p6
6.1 Data Specialists	
6.2 PI Recognition	
6.3 Funding for Data Management	
7. Cruises and Projects - Minimum metadata requirements.....	p8
7.1 Metadata	
7.2 Metadata Directory Interchange Format (DIF)	
7.3 Metadata Cruise Summary Report (CSR)	
8. Data Centres - a possible model.....	p9
7.1 National and specialist data Centres	
7.2 Validation and quality control	
7.3 Archiving	
9. Timescales for data delivery, sharing and release.....	p10
9.1 Timescales for data submission	
9.2 Timescale for data release	
9.3 Formal statement of data sharing policy for IMBER	
10. Next steps.....	p10
11. Summary of Recommendations.....	p14

## List of appendix

- Appendix 1 List of attendees
- Appendix 2 Detailed list of data types
- Appendix 3 Individual presentations
- Appendix 4 Abbreviations and definitions
- Appendix 5 Terms of reference for the DMC

## Data Management in IMBER

*Italics means not updated from GEOTRACES text*  
([http://www.ldeo.columbia.edu/res/pi/geotraces/documents/geotr\\_DataMant4SSC\\_final\\_000.pdf](http://www.ldeo.columbia.edu/res/pi/geotraces/documents/geotr_DataMant4SSC_final_000.pdf)).

### Report from Data Management Committee June 2007

#### 1. Executive summary

*Recognizing the importance of data management for the IMBER project, the IMBER Planning Committee convened as one of its first activities a meeting designed to launch IMBER data management. Initiation of data management is a crucial prerequisite for IMBER field research. The meeting described in this report resulted in recommendations for a IMBER data management system and data policies to guide IMBER scientists. The data management system should include a Data Management Committee, a Data Liaison Officer in the IMBER International Project Office, Data Specialists on IMBER cruises and associated with process studies, and two Data Assembly Centres that will manage IMBER data. Proposals are also put forward for metadata and data collection; time scales for data and metadata delivery to data centres and participants; timescales for public release of data; and a data sharing policy. Appendices set out much of the background thinking on which these proposals are based. This report will be delivered to the IMBER Scientific Steering Committee for consideration and action at their first meeting, in mid-2006.*

#### 2. Preamble

*In order to propose a data management system it is first necessary to have a good idea of the data types, their characteristics and quantities that must be handled. The meeting therefore began with an overview of IMBER. The details of data management—from collection to final archiving—were then discussed. This report follows the same structure. Much of the detailed material and presentations will be found in appendices, so that the main conclusions and data policy proposals are relatively short. A significant number of the recommendations have been lifted, with little or no change, from previous data management meetings or programmes (see <http://www.jhu.edu/scor/DataMgmt.htm>). Meeting participants recommended adopting, as much as possible, successful approaches used by other projects.*

#### 3. Objectives of the first DMC meeting

- Detail the data types and data management requirements of IMBER
- Review the experiences of previous projects and how these can contribute to forming a IMBER data management system
- Specify data management policies for IMBER
- Design a IMBER data management system, or set out and compare alternatives

- Document the next steps and the time scale on which progress is desirable
- Create a report for consideration by the IMBER Scientific Steering Committee

#### **4. IMBER goals and themes**

IMBER is a decade-long international programme (2004-2014) that develops new knowledge of ocean biogeochemical cycles and ecosystems. IMBER is chaired by Julie Hall and vice-chaired by Patrick Monfray and Dennis Hansell. This program is co-sponsored by the International Biosphere-Geosphere Program (IGBP) and the Scientific Committee on Oceanic Research (SCOR).

IMBER's challenge is to provide a comprehensive understanding of and accurate predictive capacity for, ocean responses to accelerating global change and the consequent effects on the Earth System and human society. The goal is to investigate the sensitivity of marine biogeochemical cycles and ecosystems to global change, on time scales ranging from years to decades. To achieve this goal, the Science Plan for IMBER is organised around 4 major research themes.

- Theme 1: Interaction between biogeochemical cycles and marine food webs.

The goal is to determine what are the key marine biogeochemical cycles, ecosystem processes, and their interactions, that will be impacted by global Change?

- Theme 2: Sensitivity to global Change, (Theme 2 is the core of IMBER).

The goal is to determine what are the responses of key marine biogeochemical cycles, ecosystems and their interactions to global change?

- Theme 3: Feedbacks to the Earth System

The main question is what is the role of ocean biogeochemistry and ecosystems in regulating climate?

- Theme 4: Responses from Society.

Basically four kind of projects have been defined: "endorsed projects", "regional activities", "national activities" and "contributing projects" (such as EUR-OCEANS and CARBO-OCEANS). IMBER Science encompasses different types of studies (long term observations, hydrographic lines, process studies, mesocosm experiments, lab experiments, models) and is multidisciplinary (biogeochemistry, chemistry, biology, physics, hydrography).

#### **5. IMBER data management**

##### **5.1 Introduction**

A number of presentations made during the meeting are included or summarized in Appendix 3. Much reference was also made to lessons learnt from previous projects, which are well documented at [www.jhu.edu/scor/DataMgmt.htm](http://www.jhu.edu/scor/DataMgmt.htm). Here we present for the IMBER SSC our proposals on Data Policy and Procedures for IMBER.

We shall distinguish between "metadata," which describe a data set, from the "data" themselves. Metadata comprise the essential information about how, what, where, when and by whom data were produced. Without them, the actual data are worthless. Both metadata and data are in the DMC's remit.

## 5.2 IMBER data

IMBER endorsed research projects will, by definition, be multidisciplinary in nature. IMBER will not be able to proscribe standardized methodologies that will allow for direct comparison of individual measurements or the collation of global datasets composed of biogeochemical and ecosystem parameters. Rather, the expectation is that comprehensive, regionally specific collections of sustained observations and process studies will be constructed. IMBERs' legacy will be a multidisciplinary distributed dataset that is publicized through an online public access portal. Because of the variety of data types, IMBER will make use and build upon existing data centers from different marine disciplines, but centralize the access to data. This will facilitate scientists that conduct IMBER relevant research to intercompare and synthesize data from different projects and from different geographical regions. Results from individual projects and the use of the data product for modelling efforts will further IMBER's goal of quantifying the sensitivity of marine biogeochemical cycles and ecosystems to accelerating global change

## 5.3 Oversight of IMBER Data Management

The IMBER SSC has appointed a Data Management Committee (DMC) comprising data originators (i.e., observational scientists), data managers (national and international) and data users (including modellers). The DMC is assisted by a Data Liaison Officer in the IMBER International Project Office (IPO). There need to be, in addition, national and specialist data centres (as discussed below), which we shall refer to as IMBER Data Centres. Data Specialists (6.1) should support projects and cruises.

The IMBER DMC has three areas of responsibility (taken and modified from <http://www.jhu.edu/scor/DMReport.pdf>):

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

The terms of reference for the DMC (Appendix 5) are derived from these areas of responsibility, augmented by items from other major projects.

### Role of IPO in Data Management

We agree with the recommendation of the SCOR/IGBP meeting on data management (<http://www.jhu.edu/scor/DMReport.pdf>) that management of metadata is best handled by a person working at the IPO. This may not be a full-time job, but the time commitment should not be underestimated. A Data Liaison Officer (DLO) should be employed at the IPO from an early stage, with the following responsibilities

- Maintain a list of IMBER cruises

- Keep track of project metadata
- Work with each IMBER project to encourage development of their data management system.
- Maintain a catalogue of actual and expected data sets by producing Directory Interchange Format (DIF) or equivalent discovery metadata records (conforming to the ISO19115 standard for metadata)
- Ensure that standardized parameter descriptions are adopted in DIF records (as will be developed for IMBER by the DMC)
- Ensure that names of cruises, station positions, etc are unique
- Interact with IMBER Data Centres to coordinate their activities and interactions with PIs
- In particular, ensure timely delivery of metadata and actual data to the IMBER Data Centres
- Contribute to and maintain the project web-site

The DLO will be an ex officio member of the DMC, and will report to the Director of the IPO and to the DMC.

## 6. End to end data management - role of Data Specialists

Managing data involves a complex set of operations, including collection, calibration, documentation, submission to data centres, quality control, dissemination and archival. Scientists are not usually very good at managing this entire set of important operations (apart from their prime responsibility to collect top-quality data), and it is strongly recommended that data management professionals be involved in all IMBER data activities from the start. We shall call this “end-to-end” data management. In essence, end-to-end data management means involving a data centre from the planning stage, including participation on cruises or projects and involvement in data collection, as we now describe.

Responsibility for **data collection** is spread among many individuals – PIs, their technicians, students, etc. But for a project (or on a cruise) it is the Principal Investigator's (Principal Scientist's) responsibility to ensure that metadata and data documentation are completed and delivered to the IPO. This is a major task, and it is strongly recommended that a person with data management experience be appointed, delegated or hired to serve as the project Data Specialist (DS). Allocation of funding to Data Specialists will pay off amply in the completeness and quality of the IMBER final data sets.

### 6.1 Data Specialists

The Data Specialists should be tasked to help the PI (Principal Scientist) with data issues, and would be briefed before the start of the project (cruise) by the Data Liaison Officer or other experienced data manager. Responsibilities of the Data Specialist will include

- Ensure that suitable log sheets have been provided for all activities

- Assist and support scientists in preparation of metadata
- Maintain regular checks that all logs are being correctly completed
- Assemble all metadata from the project (cruise)
- Assist with preparation of data files, ensuring that all necessary parameters are included
- Evaluate the quality of data, either by personal expertise or by discussion with PIs, and help to document quality and missing or suspect data
- Facilitate assembly of data sets and make data integrity checks.

This list is far from complete, and should be expanded by the PI as required. The duties of the Data Specialists will vary from project to project (cruise to cruise) depending on their personal expertise and the details of each activity, but some items will always be included, primarily concerning collection of complete metadata. Data Specialists may be employed by the PI, Principal Scientist or another cruise scientist as part of a grant, employed by the ship-operating institutions or employed by an IMBER Data Centre.

## 6.2 PI Recognition

To meet project goals for understanding processes in the ocean, and create a long-term legacy in terms of data produced by the project, the project's data must be submitted to a data center and made available to other scientists. Encouraging project scientists to submit their data to a recognized database requires that data management systems provide benefits for data submission. Project data management systems can encourage data submissions by actions such as backing up PIs' data at an early stage, providing security against data loss, helping with calibration and validation, serving as a long-term archive, and answering requests for data. One important way to encourage data submission is to provide principal investigators recognition for data submissions, and not just punishing infractions. One form of recognition that is used widely in other fields, such as molecular biology, is to assign a persistent identifier for data and a venue for publishing the data, so that the data publication can be cited like a scientific paper. Two marine science journals already either require or encourage data submission: *Marine Micropaleontology*<sup>1</sup> and *Geochemistry, Geophysics, Geosystems*<sup>2</sup>. IMBER scientists should be encouraged to submit their data to a recognized national or international database in conjunction with publication of IMBER-related research papers. Additionally, IMBER should support the efforts of SCOR to investigate the use of digital object identifiers and other persistent identifiers to make project data sets citable.

### References

Bindoff, N, and D. Legler. 2003. The WOCE global data resource: Lessons for CLIVAR data and data requirements. *CLIVAR Exchanges* No. 26, pp. 1-6.

---

<sup>1</sup> See [http://www.elsevier.com/wps/find/journaldescription.cws\\_home/503351/authorinstructions](http://www.elsevier.com/wps/find/journaldescription.cws_home/503351/authorinstructions) "If the original data in the submitted manuscript are not available at an internationally recognized electronic database, they must be submitted as tables or as appendices; the latter will be published electronically only. If the data are available on-line, please provide the url."

<sup>2</sup> This journal already provides the possibility for publishing "data briefs" that "report previously unpublished data, with appropriate documentation, accompanied by a minimum of interpretation and discussion" (see <http://www.agu.org/journals/gc/>).

### **6.3 Funding for Data Management**

Planning adequate funding for data management is a vital issue for each research component of IMBER to face early in its development. Estimates for similar large international research projects indicate that 6-10% of a research project's budget should be devoted to managing its data (Bindoff and Legler, 2003; Glover et al., 2006). National funding agencies should ensure that such funding levels are available and require that the projects they sponsor provide well thought out data management plans for their research data. Funding should be available from the beginning of the projects and carried through synthesis at their conclusion. Failure to support project data management adequately, both in terms of personnel and finances, reduces the impact of a project and its legacy.

## **7. Cruises and Projects - minimum metadata requirements**

### **7.1 Metadata**

Endorsement of a scientific activity by IMBER requires metadata to be submitted on the shortest possible time scales. Discovery metadata (what was collected where, when and by whom) should be submitted by project scientists with the help of the Data Specialist to the Data Liaison Officer at the IPO as the cruise is planned and when the research vessel docks at the end of a cruise. Failure to do so should be considered reason to remove IMBER endorsement, as the lack of access to metadata compromises the ability of IMBER to fulfil its goals. Involvement in IMBER should also entail a credible commitment to the timely submission of data to a project-approved database to ensure long-term archiving of the data.

### **7.2 Metadata Directory Interchange Format (DIF)**

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. It is recommended that IMBER should adopt this standard to document the data collected during all IMBER cruises and other activities (mesocosm studies, lab-based experiments, mooring and/or coastal time-series sampling, etc.). DIFs will be searchable via GCMD and IMBER could request GCMD to set up a customized IMBER portal for the IMBER DIFs. Alternatively or additionally, it should be possible for IMBER's DLO to develop IMBER's own customized interface to manage the DIFs if felt necessary.

### **7.3 Metadata Cruise Summary Report (CSR)**

Cruise reporting requires a specific set of metadata information. Cruise Summary Reports have been developed by ICES and widely adopted by the oceanographic community worldwide as a standard for cruise reporting. It is therefore desirable that IMBER recommend the use of CSRs for reporting metadata related to IMBER cruises.

CSRs will need to be filled in at the end of each IMBER cruises. This should be done by the PSO or the appointed data specialist. CSRs should then be submitted to their NODCs if it is already an established protocol. The NODC will then ensure that the CSR are forwarded to the international CSR database. If CSRs are not managed by the

NODC or if they are no NODC in the country then CSR should be submitted directly to the international CSR database via the SeaDataNet web site. Submission of CSR should be done as soon as possible after the cruise has ended.

Cruise Summary Reports will need to be converted to DIF. It is envisaged that this will be the responsibility of the DLO. It might be needed to develop a script that convert CSR to DIF (who?). In order to do this a suitable mapping between CSR's parameter categories and GCMD parameter valids will need to be built up. Roy Lowry has offered to create such mapping.

## **8. Data Centres - a possible model**

### **8.1 National and Specialist Data Centres**

IMBER should work with one of the Earth sciences data centres to develop an IMBER data portal, providing integrated access to IMBER data through the Internet. Access through the portal should allow for flexible, ad hoc queries and data downloads to common formats; and both public and private access, which should be independent of the location of the data. A security layer is needed to allow access to non-public data for PIs with permission, including capability for groups of PIs to access subsets of the data, based on approved user lists and passwords. Permissions will have to be managed by the DLO, according to policies developed by the DMC and approved by the SSC. The portal, and its underlying cyber-infrastructure, should also adhere to IMBER data policies regarding data availability, proprietorship and release.

The DLO should maintain a catalogue of all data available, including metadata types. The portal should make the catalogue, with contact information, easily available to facilitate data access. The catalogue will encourage collaboration within IMBER, in addition to timely linkages with other research programs (e.g., IMBER, SOLAS, GEOHAB, LOICZ).

The portal should be supported by a relational database housing the IMBER water column data and any of the IMBER data streams that are not housed at other already-existing data storage and access facilities. Data streams that are housed at other data centres (e.g. CCHDO, see below), along with their relevant metadata, should also be accessible through the portal. The database should be structured so that metadata can be accessed through the portal, either separately from or along with the data. A metadata report should be available that offers users a list of the metadata available for each of the data categories. Data downloads should always include at least the version, units, quality and citation metadata for each dataset. Preliminary data, not yet publicly released, should be accompanied with a message clearly stating as much.

The IMBER DAC should provide data support services for nations or smaller IMBER programs that lack their own data support infrastructure. Support would be required from the DAC for formatting and transmission of measurements and metadata.

The IMBER data portal should accept and encourage feedback from users regarding technical and data quality issues. The DAC should be responsible for technical access issues. Issues related to data quality should be forwarded by the DAC to the relevant PI,

with a copy to the IMBER DLO. The IPO should track that data quality issues are addressed by the PIs.

Water column data that are compatible with the CLIVAR hydrographic data types should be submitted to the CCHDO at Scripps. The water column data are the central data set of IMBER, so IMBER will benefit greatly from the expertise of CCHDO. In addition, many of the IMBER hydrographic data sets are likely to be part of a continuum of data sets from other programs, e.g. CLIVAR, WOCE etc. Thus it is important that these aspects of the IMBER data sets be quality controlled, stored and archived in a seamless manner with the other global data sets. However, the CCHDO does not have the capability to handle the wide range of data types that IMBER will collect, so cannot be the overarching IMBER DAC. The CCHDO also does not hold its data in relational databases, which IMBER would like to make use of to facilitate data integration across cruises and between distinct data types. Thus, an overarching data centre must be developed. The water column data submitted to the CCHDO will be accessible (for example by mirroring) in the IMBER database.

## **8.2 Validation and quality control**

Individual PIs are responsible for quality control of their data and should provide in their metadata notes about any doubtful data values. Questionable data should be flagged rather than discarded. Participants using the data should report questionable data to the DACs (who also will be noting problems), or possibly to the PI copied to the DAC. The DACs will pass data quality problems back to the PIs. The cruise or process study Data Specialist and the Data Liaison Officer should be copied into such correspondence and may be able to help, at least in the documentation.

Experience shows that data quality problems are often revealed once the data begin to be used scientifically, often by individuals other than the data collector. Comparison with other parameters or comparison between data sets at the same location can reveal errors or offsets. This is one good reason for making data available to other participants without delay. More formally, groups of PIs expert in a particular parameter (for example Fe) will be encouraged to apply further quality controls, such as cruise intercalibration.

## **8.3 Archiving**

The data from IMBER must ultimately be preserved for posterity. In part this can be done by providing data sets in multiple copies (such as DVDs), but all media degrade on some time scale and the formats and technical specifications of storage devices are constantly evolving. The World Data Centres (WDCs) have the resources to preserve data for the long term and periodically to update the media on which data are held. Meeting participants believe that the World Data Centre for Oceanography, Silver Spring would be the most appropriate data centre to hold most IMBER data. Personnel from this WDC should be encouraged to interact with the DMC from an early time to facilitate eventual long-term data archiving at the completion of the programme.

## 9. Timescales for data delivery, sharing and release

### 9.1 Timescales for data submission

The IPO should take the lead role in maintaining a catalogue of project metadata. Operating a project metadata catalogue should be considered a core activity of the IPO and will be a major duty of the Data Liaison Officer. The rationale is as follows.

Both metadata and data need to be submitted to the Data Liaison Officer and the relevant DAC as soon as they are created; metadata about cruises (when, where, who, what will be measured) should be submitted when a proposal is funded and cruise metadata should be submitted immediately after the end of the cruise. The most important reason for this is data security - with the best will in the world individual PIs may lose metadata or even the data themselves and duplication is extremely important to provide backup. Metadata comprise, in part, the detailed lists being made during the cruise or process study and the Data Specialist (working with the lead scientist) should ensure that these lists are printed, copied or scanned on board the ship (or during the process study) and that duplicates are delivered to the DLO immediately.

Cruise reports should be submitted to the DLO within 6 months of the end of the cruise. The cruise reports will be publicly available, distributed through the IMBER portal and, if possible, be assigned a reference digital object identifier (DOI).

The data assembled on board ships should be delivered to the appropriate DAC within one month of the end of the cruise (for exceptions see below). These may be preliminary or partially calibrated data, as accompanying metadata will show, and the most important reasons for such rapid delivery is to start the dialogue with the DAC on the quality and completeness of the data as well as for data duplication and security. Participation in a cruise implies willingness to share data with other cruise participants. The DAC will be able to check completeness with the relevant PI and may be able to help with error checking. It is recognised that, for many data, "one month" delivery is impossible, and a table of expected delivery times will be developed to allow for necessary delays for particular parameters, for example, isotope ingrowth.

It is expected that cruise participants will have access to routine hydrographic parameters as soon as they are available, either shipboard or as soon as available at the DAC. Whether this access extends to the individual TEI data sets of other PIs is an issue that needs discussion and resolution by the IMBER SSC - there are arguments for and against this approach.

The major incentive for rapid data submission is the extra support that will then be available to quality control and to interpret the originator's data. The data will be integrated with the authoritative metadata without personal effort. The DAC will quality control data and may help the originator in detecting problems. Comparison with the data sets of other PIs is one sure way to tease out errors and inconsistencies. Access to related data will aid in linking of individual data sets and promoting scientific collaboration. Data will be professionally maintained, safeguarded and archived.

#### Timescales for data submission - summary:

Metadata - as soon as created from the planning stage onwards

Data (not finalized) - within 1 month (of collection, or end of cruise), with possible approved extension as tabulated

Cruise report - within 6 months of the end of the cruise

Final detailed data report for each process study (see section 5.6) - within 6 months of the end of the process study

Final data - within 2 years, exceptions possible from the IMBER SSC

## **9.2 Timescales for data release**

IMBER must adopt the ICSU principle of free and open data exchange, which will help IMBER achieve its goals. Data release can be split into two categories: (1) release to other participants in the cruise and (2) public release. The DAC will have to hold a list of “participants” (which may include closely associated persons who were not actually at sea, as agreed with the Principal Scientist) for each cruise. In general, data should be available to participants without delay, but the DAC will need to consult the relevant PI or keep blanket permission information with the metadata. Procedures for restricted access to data are well established (e.g., using paired passwords and approved user lists) and will be implemented by the DAC(s). Participants in other IMBER cruises should contact the relevant PI directly to obtain desired data; they will be aware of the existence of the data through the metadata maintained by the IPO. Public release will normally be two years from the end of the cruise or field activity, but should be extended when analytical procedures have inherent built-in delays.

### Timescales for data release - summary:

Participants in a particular cruise - as soon as available at the DAC with knowledge and permission of the relevant PI

Public release - within two years of end of cruise (+ extra time for particular data type as approved by SSC)

## **9.3 Formal statement of data sharing policy for IMBER**

In light of the above discussion, the data sharing policy is as follows:

There is a fundamental trade-off in IMBER - on the one hand, protection of the intellectual effort and time of originating investigators (those who plan an experiment, collect, calibrate, and process a data set to answer some questions about the ocean), and on the other hand, the need to compare various data sets and data types to check their consistency, to better understand the ocean processes involved, and to see how well the numerical models describe the real ocean. The policy adopted by IMBER is a trade-off between these conflicting needs.

IMBER activities require all participating scientists to submit their metadata to the Data Liaison Officer at the IPO as soon as they are available (including cruise/process study details when the proposal is accepted for funding). Any metadata and data produced during the cruise/process study should be made available to participating scientists immediately in preliminary form during the cruise/process study. “Routine data,” by which we mean the basic hydrographic parameters, will be made public a short time after each cruise/process study is completed. Preliminary data collected as part of IMBER are to be submitted to the DAC within a month of their collection for the purposes

of quality control and data synthesis during the 2-year “publication rights period.” Any data collected as part of IMBER should be made publicly available no later than 2 years from collection, with an extension of this period as specifically granted by the IMBER Scientific Steering Committee (SSC) for particular parameters that require extensive processing after the cruise or process study is completed, and recognising that some nations do not permit release of data collected within their EEZs. Prior to public release, all data will be considered preliminary. Such data will be available to participating scientists, who should consult the data originator about its status. Data should be shared with other cruise/process study participants as soon as they become available during or after a cruise or process study, to enable data synthesis to proceed rapidly, with the understanding that the data are the proprietary material of the originating scientist and may not be used without their permission. However, for non-participating scientists the data can be obtained only with the permission of the responsible participating scientist.

The recipient DAC will not publicly redistribute such data, or a derivative containing most of the information during the publication rights period.

The receiving investigator should not publish any paper based predominantly on the received data during the publication rights period, should co-author results with the originating investigator, and should not redistribute the data.

Adherence to this data policy is expected of all scientists participating in national and international IMBER activities.

### ***Special requirements of process studies***

*A simple, convenient data policy and formatting system is the best way to ensure international agreement and will facilitate and promote regional cooperation and collaboration. Nations without oceanographic data centres and/or lacking funds to build and maintain data management facilities will then be encouraged to submit their data sets to the international data centre.*

- *Regardless of data storage details, the process study results should be distributed through the same portal as the core activity data.*
- *End user data retrieval for process study results should have the same interface (look and feel) as core data.*
- *Process study results should use the same data stream processing facilities as used for core activities whenever possible.*
- *Established data processing facilities should be used rather than building new capabilities whenever possible.*
- *Process study metadata (who, what, when, where, how) should be submitted to the IMBER data centre as soon as soon as possible (within one month after completion of the field work or each phase of the field work) using either a detailed preliminary project report or standardized metadata forms specifically designed by the data manager for the specific process study.*
- *A final detailed data report is required for each process study. These reports will be similar to a cruise report. In addition to describing the process study, these reports will provide one mechanism to give credit to data generators. The final*

*reports will be publicly available, distributed through the IMBER portal and, if possible, be assigned a reference DOI.*

## **10. Next steps**

Sophie to contact each IMBER project to encourage development of their data management policy

Sophie to develop the IMBER web pages on data management

Sophie to print off list of projects for the SSC

Specialist data centres - letter to CCHDO from SSC

Roy and Gwen to create CSR to DIF converter

Sophie to contact GCMD

Todd to create IMBER portal to GCMD

Todd to create a Template for DIFs

Gwen to set up an initial list of agreed terminology for IMBER DIFs (GCMD DIFs)

Capacity building and training: all to organise DM trainings when possible

### Issues for SSC

Funding for travel of DMC members, Sophie training.

Fund some travel to IMBER open science meeting

National data managers meeting = data specialist training.

50% from SCOR of DLOs time

## **11. Summary of Recommendations**

1. IMBER should adopt, as much as possible, successful approaches used by other projects.
2. IMBER should establish a Data Management Committee (DMC) comprised of data originators (i.e., observational scientists), data managers (national and international) and data users (including modellers).
3. A Data Liaison Officer (DLO) should be appointed to work at the IMBER International Project Office (IPO). The primary responsibility of the DLO will be to define and maintain metadata for IMBER. The DLO should be appointed as soon as the IPO is established.
4. The IMBER SSC should establish the DMC and the IPO and appoint the DLO in 2006, and begin to develop relationships with existing and potential DACs in 2006. All of these require that funding be sought.
5. IMBER should make use of the CLIVAR and Carbon Hydrographic Data Office (CCHDO), which is already funded to take in profile data internationally, but

should arrange another DAC to develop a relational database to serve all IMBER data.

6. IMBER should work with one of the Earth sciences data centres to develop a IMBER data portal, providing integrated access to IMBER data through the Internet.
7. It is strongly recommended that data management professionals be involved in all IMBER data activities from the start. This is known as co-operative or “end-to-end” data management. Practically, this should be achieved by the assignment of a Data Specialist to each cruise or process study (in addition to the DLO at the IPO), whose primary role is to assist the lead scientist in metadata collection and data management. Allocation of funding (from a PI’s grant or a Data Centre) will pay off amply in the completeness and quality of the IMBER final data sets.
8. Endorsement of scientific activity by IMBER requires metadata to be submitted on the shortest possible time scales. Failure to do so should be considered reason to remove IMBER endorsement.
9. IMBER should adopt the Directory Interchange Format or a similar discovery metadata standard, for example, ISO19115 when it is completed.
10. Metadata should be delivered to the IPO as soon as created, from the planning stage onwards. Data (not finalized) should be submitted to a DAC within 1 month (of collection, or end of cruise), with possible approved extension as tabulated. Cruise or project reports should be submitted to the IPO within 6 months of the end of the cruise or process study. Final data, following international guidelines, should be submitted within 2 years of cruise or process study completion, with exceptions possible from the IMBER SSC.
11. Data should be shared with other participants in a particular cruise or process study from an early stage (as soon as it available at the DAC, with knowledge and permission of the relevant PI)
12. Public release will normally be two years from the end of the cruise or field activity, but should be extended when analytical procedures have inherent built-in delays.
13. Individual PIs are responsible for quality control of their data. Groups of PIs expert in a particular parameter will be encouraged to apply further quality controls.
14. The IMBER DMC, when approved, should discuss specific data management requirements for IMBER process studies, when it is better known what these studies will be.
15. The IMBER SSC should consider providing access to project-related publications through a publication database, such as that used by GLOBEC.
16. The World Data Center for Oceanography, Silver Spring, should be considered as the most appropriate WDC for IMBER data. It is recommended that the DMC contact this WDC to tell them about IMBER once the project is under way, and discuss long-term archiving.

## Appendix 1 – list of attendees

Dr Sophie Beauvais  
IMBER International Project Office,  
Institut Universitaire Européen de la Mer (IUEM),  
Technopôle Brest-Iroise, Place Nicolas Copernic  
29280 Plouzané, FRANCE  
Email: [sophie.beauvais@univ-brest.fr](mailto:sophie.beauvais@univ-brest.fr)  
Phone: +33 2 9849 8693

Dr Jay Cullen  
School of Earth and Ocean Sciences  
University of Victoria  
P O Box 3055 STN CSC  
Victoria, BC, CANADA  
[jcullen@uvic.ca](mailto:jcullen@uvic.ca)  
Phone: (250) 472 4353

Dr Julie Hall  
Group Manager Aquatic Ecology and Ecotoxicology  
National Institute of Water and Atmosphere (NIWA)  
PO Box 11 115, Hamilton  
NEW ZEALAND  
[j.hall@niwa.co.nz](mailto:j.hall@niwa.co.nz)  
Phone: +64 7 856 1709

Dr Wilco Hazeleger  
KNMI, P O Box 201,  
3730 AE De Bilt  
THE NETHERLANDS  
[Wilco.Hazeleger@knmi.nl](mailto:Wilco.Hazeleger@knmi.nl)  
Phone: +31 30 2206718

Dr Gwenaëlle Moncoiffe  
British Oceanographic Data Centre  
Joseph Proudman Building  
6 Brownlow Street, Liverpool L3 5DA, UK  
[gmon@BODC.ac.uk](mailto:gmon@BODC.ac.uk)  
Phone: +44 (0)151 7954880

Dr Todd O'Brien  
National Marine Fisheries Service,  
1315 East-West Hwy,  
Silver Springs MD 20910, USA  
[Todd.Obrien@noaa.gov](mailto:Todd.Obrien@noaa.gov)

Prof. Raymond Pollard

National Oceanography Centre, Southampton  
European Way, Southampton SO14 3ZH, UK  
[rtp@noc.soton.ac.uk](mailto:rtp@noc.soton.ac.uk)  
Phone: +44 (0) 23 80596433

Prof. Reiner Schlitzer  
Columbusstrasse  
D-27568 Bremerhaven (Building D-1160)  
GERMANY  
[rschlitzer@awi-bremerhaven.de](mailto:rschlitzer@awi-bremerhaven.de)  
Phone: +49 (471) 4831 1559

Dr Toru Suzuki  
Marine Information Research Centre  
[Suzuki@mirc.jha.jp](mailto:Suzuki@mirc.jha.jp)

Dr Ed Urban  
SCOR Secretariat, Dept of Earth and Planetary Sciences  
The Johns Hopkins University  
Baltimore, MD 21218, USA  
[Ed.Urban@jhu.edu](mailto:Ed.Urban@jhu.edu)  
Phone: +1 410 516 4239

## **Appendix 2 – detailed list of data types**

- Sustained long term observations
- Repeat hydrographic lines and basin scale transects
- Field-based process studies
- *In situ* mesocosm experiments
- Field and laboratory based experiments
- Use of paleo-proxies
- Models

## **Appendix 3 – individual presentations**

### **A3.1 Data management overview** by Roy Lowry

For definitions of data management technical terms, see Appendix 4.

#### Data Management Issues

Data management issues include project dataset scope, description, instantiation (for definition see Appendix 4), and quality assurance; interoperability of databases and data standards; project data policy; intra-project data exchange; data management infrastructure; and long-term stewardship of the data.

### Project Dataset Scope

The project dataset is an aggregation of datasets from different project activities. Building a catalogue of these activity datasets is the most basic task of project data management. However, it can be far from straightforward. There can be controversy about what comprises a project cruise. For example, Belgica OMEX cruises were considered as totally unrelated to JGOFS by some people and totally contributory to JGOFS by others. Scientific Steering Committees have trouble saying no to willing contributors of data and have a hard time constraining the project scope.

There can be over-enthusiasm at the national level for activities to include in the project dataset, even if peripheral to the project objectives (one nation submitted several hundred cruises to the JGOFS database, most of which were of peripheral relevance). Some project activities are shared cruises with other projects, creating data ownership and access issues. For example, who should have access to CTD datasets from WOCE cruises carrying a JGOFS team doing carbonate system measurements? It is imperative that such issues are addressed by the responsible SSCs and IPOs before cruises take place.

### Project Dataset Descriptions

Dataset catalogues need to be more than just a list of dataset names. Each dataset needs to be described by a discovery metadata record, which provides data that makes it possible to find the data based on relevant descriptors. According to the 2003 SCOR/IGBP data management meeting (<http://www.jhu.edu/scor/DataMgmt.htm>), discovery metadata should be compiled and managed by the IPO.

### Project Dataset Instantiation

A collection of data interchange formats (DIFs) needs to be converted into a collection of data files that make up the datasets described. This process requires careful planning and management during the life of the project.

### Project Dataset Quality Assurance

Every project must determine who is responsible for project data quality.

PIs can be given responsibility for specialist parameters. But who ensures that all CTD parameters are fully worked up and calibrated? And who is responsible for metadata quality assurance? These issues must be worked out by the IMBER Data Management Committee.

### Interoperability and standards

Data standards are necessary to ensure interoperability of databases. WOCE had strong syntactic and some degree of semantic data standards, but JGOFS did not. As a result, there was an integrated WOCE dataset available at the end of the project, but an integrated JGOFS dataset needed 5 years of post-project work to produce.

The value of data and metadata interoperability cannot be overstated. Interoperability can be achieved most easily through universal adoption of standards. Syntactic interoperability comes easily through adoption of mature technologies (e.g., netCDF). Semantic interoperability is much harder to achieve. Content standards and controlling vocabularies for soft-typed elements are essential to achieve semantic interoperability.

### Data policy

A project data policy specifies who can have access to what data and when. The policy specifies who gets what reward for what actions. Simple policies are the simplest to implement. Liberal policies also are simple to implement, but may not offer the desired features. The project data policy needs to be specified and agreed at the outset. As part of the policy, the issue of data release to the public domain needs to be addressed

The data policy also needs to specify how project participants obtain project data held by other participants. Today, anything other than Web-based data access is inconceivable. The intra-project data exchange expected within IMBER has infrastructure implications.

Major functions of a data policy are expectation management and policing the consensus.

### Infrastructure

The design of the project data infrastructure has many issues that should be considered:

- Will the system have visualisation/usage tools? The answer to this question may influence other decisions, such as standards. One of the valuable features of the CLIVAR data management system on cruises is the availability of data-plotting software that combines cruise data in near-real time.
- To what extent will the project data management system be a distributed system versus a centralised system? The desirable granularity is the crucial issue. Too fine granularity (e.g., each PI's computer as a node), results in poor reliability. A coarser system requires ingestion of data from distributed data origination sites and therefore resources to accomplish this integration. A single centralised system may be unrealistic, due to the resources required.
- Will a DAC concept be used for designated data types? This approach worked well for WOCE. DACs for specific data types can be shared with other projects, which is an important consideration as other new marine research projects are developing their data management systems.

### Long-term stewardship

Projects must consider long-term stewardship of project data after the project is completed and the IPO is closed. Will the data be published on physical media? Will the project maintain a Web presence on an established server after the end of the project? The JGOFS Web site has continued on the University of Bergen server for two years after the JGOFS IPO closed (so far), but it is unclear how long this university's server will host the Web site. JGOFS data DVDs (Vol. 2) will be available through WDC-MARE. For longer-term data storage, project data should be integrated into an established distributed system, specifically, one or more World Data Centres (WDCs), depending on the data type. It is very helpful for projects to involve the appropriate WDCs early in the project, rather than merely expecting to dump the data into the WDCs at the end of the project with no warning.

### A3.2 Data Centre interest in IMBER (Gwen Moncoiffe)

A	B	C	D	E	F
Specialist	Database/data centre name	Acronym	Capabilities/specialities (summarised from info provided directly or obtained/inferred from web site)	Web link	Contact
Specialist	(The) Coastal & Oceanic Plankton Ecology, Production, & Observation Database (COPEPOD)	COPEPOD	global plankton data (zooplankton, phytoplankton, ichthyoplankton, total biomass)	<a href="http://www.st.nmfs.gov/plankton">www.st.nmfs.gov/plankton</a>	Todd O'Brien, COPEPOD Project Leader (Todd.O'Brien@noaa.gov)
Specialist	Carbon Dioxide Information Analysis Centre	CDIAC	Underway pCO2 data and discrete measurements of TCO2, TALK, pCO2, and pH.	<a href="http://cdiac.esd.ornl.gov/oceans/home.html">http://cdiac.esd.ornl.gov/oceans/home.html</a>	Bob Groman ( groman@whoi.edu )
Specialist	US-GLOBEC database (hosted at WHOI)	US-GLOBEC	all data types from the US GLOBEC projects, ranging from CTD to water chemistry to satellite data to plankton nets to sea birds and mammals and also advanced technologies (video plankton recorder, ADCP, etc.)	<a href="http://www.usglobeec.org/">http://www.usglobeec.org/</a>	Edward Vanden Berghe Jeffrey H. Smart (jeff.smart@jhuapl.edu)
Specialist	Ocean Biogeographic Information System (OBIS)	OBIS	global taxonomically-defined marine biodiversity data	<a href="http://www.obis.org/">http://www.obis.org/</a>	
Specialist	Worldwide Ocean Optics Database (hosted at OMR)		a collection of several hundred ocean optics data sets gathered over time that encompass much of the world's oceans		
Specialist	Coriols and the Global Temperature and Salinity Subsurface Data Centre (hosted at SISMER)	Coriols	provides quality-controlled in-situ data in real-time and delayed modes; mainly T-S profiles and time series from profiling floats, XBT's, thermo-saliniographs, drifting and moored buoys. Gateway to global ARGO data, floats data processing centre for the European GyroScope pilot array; in situ data provider for the Mercator data assimilation project.	<a href="http://www.coriolis.eu.org/">http://www.coriolis.eu.org/</a>	
Specialist	Mediterranean Oceanic Data Base (european project)	MODB	advanced data products for oceanographic research in the Mediterranean Sea. Software products for data analysis and visualization are also prepared for distribution within the scientific community.	<a href="http://modb.oce.uva.ac.be/modb">http://modb.oce.uva.ac.be/modb</a>	
National	Japanese Oceanographic data Centre	JODC	physical (including T/S, currents, wave, sea level), bathymetry, marine pollution and marine organisms data.	<a href="http://www.jodc.go.jp/">http://www.jodc.go.jp/</a>	Tim Boyer, World Ocean Database project leader (Tim.Boyer@noaa.gov)
National	U.S. National Oceanographic Data Center (US-NODC)	US-NODC	physical and hydrographic data (T, S, O, Nutrients, Chlorophyll), produces the World Ocean Database product.	<a href="http://www.nodc.noaa.gov">www.nodc.noaa.gov</a>	Juan Brown, BODC director (jbrown@bodc.ac.uk)
National	British Oceanographic Data Centre (NERC)	BODC	all data types collected during oceanographic cruises and coastal experiments. Includes: physical, biological, and chemical discrete and continuous measurements, audio-visual data, incubation and experimental data. Partner in the SeaDataNet programme.	<a href="http://www.bodc.ac.uk">www.bodc.ac.uk</a>	
National	DEUTSCHES OZEANOGRAPHISCHES DATENZENTRUM (DSH)	DOD	oceanographic database; mainly physical and chemical variables.	<a href="http://www.bsh.de/en/Marine%20Data/Observations/DO2%20Datatabases.htm">http://www.bsh.de/en/Marine%20Data/Observations/DO2%20Datatabases.htm</a>	
National	MARINE INFORMATION SERVICE	MARIS	all data types collected during oceanographic cruises and coastal surveys. MARIS is a partner in SeaDataNet.	<a href="http://www.maris.nl/frames.asp?d">http://www.maris.nl/frames.asp?d</a>	
National	SYSTEMES D'INFORMATIONS SCIENTIFIQUES POUR LA MER (IFREMER)	SISMER	all data types collected during oceanographic cruises and coastal experiments. SISMER is a partner in SeaDataNet.	<a href="http://www.ifremer.fr/sismer/FR/donnees_FR.htm">http://www.ifremer.fr/sismer/FR/donnees_FR.htm</a>	
National	Other european data centres partners in the SeaDataNet project	SEADATANET	Pan-European infrastructure for Ocean & Marine Data Management: a network of 40 European National Data Centres working on interoperability and a common portal for data access.	<a href="http://www.seadatanet.org/ta.htm">http://www.seadatanet.org/ta.htm</a>	
ICSU World Data Centre	WDC-MARE		numeric, string, and image data related to global change in the fields of environmental oceanography, marine geology, paleoceanography, and marine biology. Uses the information system PANGAEA. Data are retrieved through the Internet via different gateways.		Sydney Levitus Vyacheslav I. Smirnov Lin Shaohua
ICSU World Data Centre	WDC for Oceanography, Silver Spring		as for US-NODC		
ICSU World Data Centre	WDC for Oceanography, Obninsk		Oceanographic data mainly physical.		
ICSU World Data Centre	WDC for Oceanography, Tianjin		physical, chemical and biological data		

#### Appendix 4 – Abbreviations and definitions

ArcIMS	software developed and marketed by ESRI to serve geographic information, such as maps, over the Web (URL). ArcIMS is a server-based product that provides a scalable framework for distributing GIS services and data over the Web ( <a href="http://www.esri.com/software/arcgis/arcims/about/overview.html">http://www.esri.com/software/arcgis/arcims/about/overview.html</a> )
BATS	Bermuda Atlantic Time-Series station
BODC	British Oceanographic Data Centre
CCHDO	CLIVAR and Carbon Hydrographic Data Office
CIESIN	Center for International Earth Science Information Network
CLIVAR	Climate Variability and Prediction project
CTD	conductivity-temperature-density measurement package
CTDO	conductivity-temperature-density-oxygen measurement package
DAAC	Distributed Active Archive Centre
DAC	Data Assembly Centre
Data orphans	Data either unforeseen due to development of new technologies or lack of mandate, awareness, and/or capability at a DAC
DIF	Directory Interchange Format
DIU	Data Information Unit
DOI	Digital Object Identifier
DLO	Data Liaison Officer
DMC	Data Management Committee
DMTT	Data Management Task Team (JGOFS)
EEZ	Exclusive Economic Zone
ENTRI	Environmental Treaties and Research Indicators
ESI Viewer	Environmental Sustainability Index Viewer
GEOHAB	Global Ecology and Oceanography of Harmful Algal Blooms programme
GISS Crop-Climate	Goddard Institute for Space Studies Crop-Climate Study
GLOBEC	Global Ocean Ecosystem Dynamics project
GCMD	Global Change Master Directory
HOT	Hawaii Ocean Time-series
HPLC	high-performance liquid chromatography
IAPSO	International Association for the Physical Sciences of the Oceans
ICES	International Council for the Exploration of the Sea
IGBP	International Geosphere-Biosphere Programme
IOC	Intergovernmental Oceanographic Commission

IMAGES	International Marine Aspects of Global Change project
IMBER	Integrated Marine Biogeochemistry and Ecosystem Research project
Instantiation	the conversion of a virtual object (such as a dataset description) into a concrete object (such as a data file holding the dataset)
IODE	Intergovernmental Oceanographic Data and Information Exchange
IPO	International Project Office
JGOFS	Joint Global Ocean Flux Study
KNOT	Kyodo North Pacific Ocean Time-series station
LDEO	Lamont-Doherty Earth Observatory
LOICZ	Land-Ocean Interactions in the Coastal Zone project
MAST	Marine Science and Technology
NERC	Natural Environment Research Council (UK)
NODC	National Oceanographic Data Centre
OAIS	Open Archival Information System
OBIS	Ocean Biogeographical Information System
ODV	Ocean Data View
OMEX	Ocean Margin Exchange project
PetDB	Petrological Database of the Ocean Floor
PI	Principal Investigator
PROOF	PROcessus biogeochimiques dans l'Océan et Flux" = Biogeochemical processes in the Ocean and Fluxes
PROVESH	Processes of Vertical Exchange in Shelf Seas (MAST)
PSDS	Process Study Data Specialist
SAN	Storage Area Network
SCOR	Scientific Committee on Oceanic Research
SDS	Shipboard Data Specialist
SedDB	Integrated Data Management for Sediment Geochemistry
Semantic data standards	standards that unify the way in which data are described by metadata, such as controlled vocabularies (lists of approved words and their definitions) and content standards
SESAR	Solid Earth Sample Registry
SIO	Scripps Institution of Oceanography
SSC	Scientific Steering Committee
SME	small to medium-sized commercial enterprise
SOLAS	Surface Ocean – Lower Atmosphere Study

Soft-typed elements items of metadata that are loosely defined by the metadata schema. For example, <temperature>25.0</temperature> is hard-typed, but <data parameter="temperature">25.0</data> is soft-typed

Syntactic data standards standards that unify the way that data are physically encoded in a file (for example, CSV or NetCDF)

TEIs trace elements and isotopes

US-Mexico DDViewer: U.S-Mexico Demographic Data Viewer

WDC World Data Centre

WDC-MARE World Data Centre for Marine Environmental Sciences

WOCE World Ocean Circulation Experiment

## **Appendix 5 Terms of reference for the DMC**

(as of June 2007)

The IMBER Data Management Committee is appointed by the IMBER SSC. The DMC comprises observationalists, modellers, data specialists and the Data Liaison Officer (DLO) from the IPO. The DLO will be an *ex officio* member of the DMC, and will report to the Director of the IPO and to the DMC.

The initial responsibilities of this committee are as follows:

- to prepare a plan for IMBER data management. The implementation plan will establish an appropriate data management strategy and policies to ensure creation of full metadata and access to, sharing of, and preservation of IMBER data;
- to develop data and metadata guidelines for IMBER related projects.

In a longer term, the DMC and the DLO shall collectively:

- ensure that the strategy is implemented;
  - ensure that IMBER related projects adhere to the policies developed by the DMC;
  - keep track of IMBER metadata and make them available to the global community, via the IMBER portal;
  - keep track of IMBER data and ensure that they are made available at first to collaborators and later to the global community;
  - to facilitate data management training for scientists.
- 
- *Oversee the work of the IMBER Data Assembly Centres (DACs) and the Data Liaison Officer in the IMBER International Project Office.*

- *Ensure that IMBER creates and maintains an integrated, international data and cruise inventory*
- *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 a suitable format (CDROM or DVD are current possibilities)*
- *Ensure that IMBER data are available for project scientific purposes and that data management meets the present scientific needs of the project without compromising future needs*
- *Monitor international acceptance of, compliance with, and adoption of, IMBER data policies*
- *Address the involvement in project data exchange activities of scientists without access to effective data management infrastructure*
- *Report regularly to and advise the IMBER Scientific Steering Committee (SSC)*