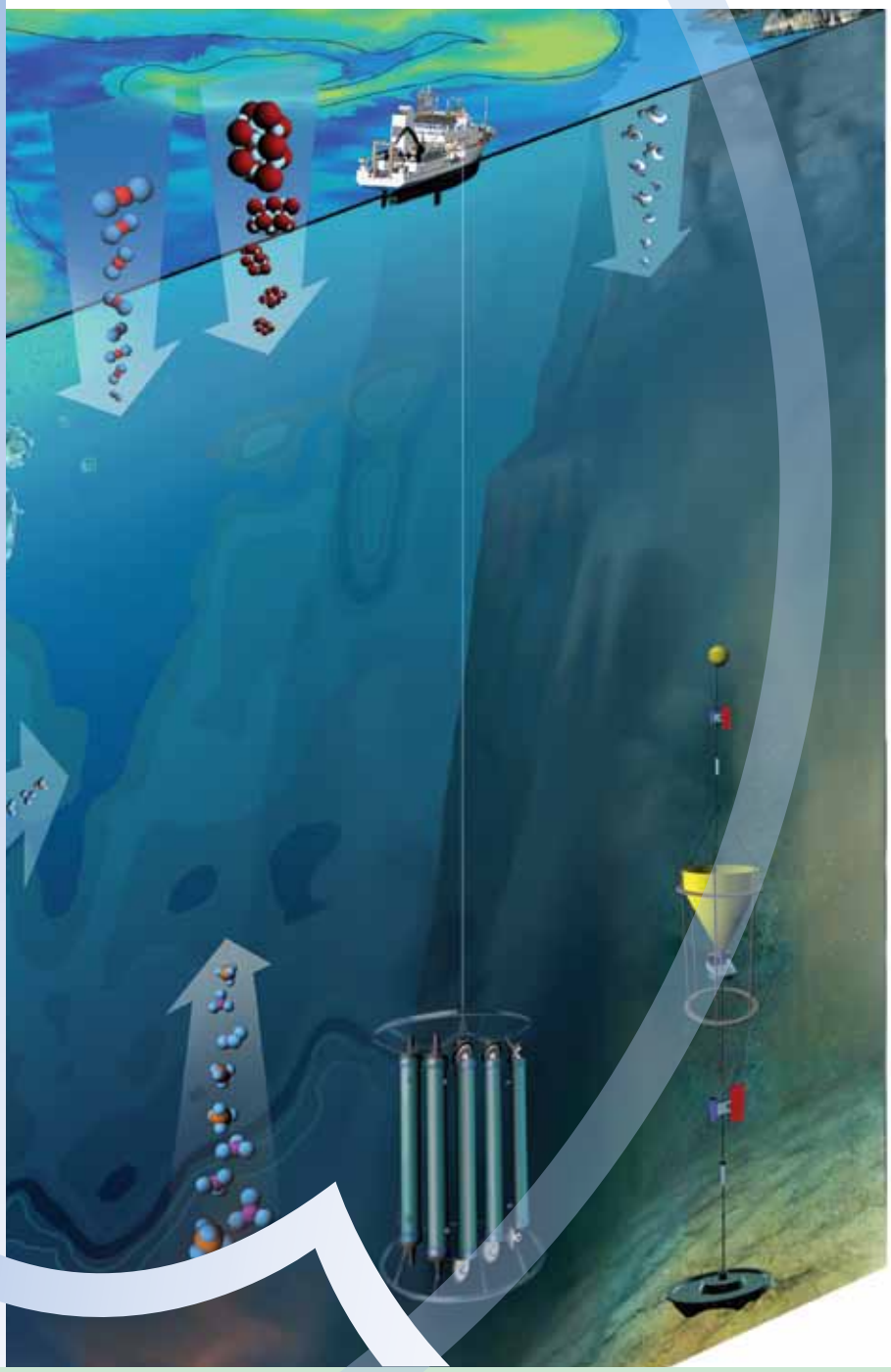
IMBER

Integrated Marine Biogeochemistry and Ecosystem Research

IMBER Report 3

The IMBER Data Management Cookbook

A Project Guide to good Data practices.



International Council for Science
SCOR
Scientific Committee on Oceanic Research

GLOBAL CHANGE
IGBP International Geosphere-Biosphere Programme

Citation

The IMBER Data Management Cookbook
- A Project Guide to good Data practices (2011).
Pollard R.T., Moncoiffé G. and O'Brien T.D.
IMBER Report No. 3, IPO Secretariat, Plouzané, France.
16pp.

Publication Details

Published by:
IMBER IPO Secretariat
Institut Universitaire Européen de la Mer (IUEM)
Technopôle Brest-Iroise
Place Nicolas Copernic
29280 Plouzané, FRANCE
Ph: +33 2 98 49 86 72

Web: www.imber.info

Graphic Designer: Hilarie Cutler

ISSN 1951-6681

The IMBER Data Management Cookbook

A Project Guide to good Data practices.

Raymond Pollard
National Oceanography Centre
Southampton, UK

Gwen Moncoiffé
British Oceanographic Data Centre
Liverpool, UK

Todd D. O'Brien
NOAA Fisheries Service
Maryland, USA

Contents

Today's Recipe:

Better Science through Better Data Management

Welcome to the IMBER Data Management "Cookbook". While the idea for this compendium of recipes to make data management digestible came from the IMBER community, the recipes are in no way restricted to IMBER, but should be suitable for any project that gathers data and wishes them to be available and useful in the long-term. IMBER is primarily a marine project, so the examples are cruise based, where a cruise involves a number of researchers, usually from different disciplines, working together on a ship at sea. The data management principles should apply to any project, even solo ones.

What's Cooking?!

Why not data management?!

Why is data management the poor relation to publication?

Why do most researchers consider that data management (DM) is the poor relation to writing papers? Because it is boring, a chore?! I'm sure you would all agree that writing papers is an essential part of a research scientist's job. Yet, once the research is done, don't you find writing the paper a chore - finding the right words, rewriting, revising, creating good figures, converting formats, responding to nitpicking reviewers, proof reading - don't you find all this time consuming and tedious? Is that any different from DM?

So why do we accept that we must write papers, but treat DM as the poor relation? Because everybody else does? That is one main reason. Because we follow the herd? YOU can change that, and this cookbook suggests how you can do that. In fact a good quality data set is a more objective legacy, as it is not biased by our interpretation. It can be reused, and it can be compared with other data sets.

Recognition for Data Management

(Go beyond "Publish or Perish"…"Data Manage or Die Morbidly"!)

Another reason why we rate DM much lower than writing papers is that we get recognition for publications. They are also often used as an indicator of professional productivity - "Publish or Perish"). They figure in our *Curriculum Vitae* (CV) and they now can be referenced in other publications using a Digital Object Identifier (DOI). There is no reason in principle why we should not also be able to reference data sets with a DOI, but the details need to be thought through. This could be a huge "carrot" to improving DM and is currently being considered by the Scientific Committee on Oceanic Research (SCOR) and the International Oceanographic Information and Data Exchange (IODE) committee of the Intergovernmental Oceanographic Commission (IOC). Examples and progress of these SCOR/IODE efforts are found in the following reports:

- Report of the SCOR/IODE Workshop on Data Publishing (Oostende, Belgium, 17-19 June 2008). IOC Workshop Report No. 207. Paris, UNESCO, 23pp. 2008. [1]

- Report of the SCOR/IODE/MBLWHOI Library Workshop on Data Publication (UNESCO Headquarters, Paris, France, 2 April 2010). IOC Workshop Report No. 230. Paris, UNESCO, 5 May 2010. [2]

Why do we need to manage data?

Proper management of your data, during and after the cruise and the project, ensures that others can use it now and in the distant future. If your data set is to be useful to anybody else, it needs to be well described, ideally while it is still fresh in your own memory. Where did you collect it? What methods? How did you calibrate? What other data are essential to using these data? Whose permission is needed to use the data? All this information is known as metadata.

What is a Data Scientist (DS)?

Just as writing a paper takes time and attention to detail, and so does data management. If you are an established researcher and the leader of a project or cruise, it is unlikely that you can spare the time to manage the data yourself, any more than you can write all of the papers that come out of the project. Our strong advice is that you seek assistance with DM, and we coin the term Data Scientist (DS) for this person who assists you. The DS may be a part-time or a full-time task, depending on the size and complexity of the project. Depending on your resources, you may be able to assign someone to the DS task, or you may have to hire someone, or you may be

1 IOC Workshop Report No. 207: http://www.scor-int.org/Publications/wr207.pdf

2 IOC Workshop Report No. 230: http://www.scor-int.org/Publications/wr230.pdf

lucky enough to be able to have a national data centre who can assign somebody to your project.

For the Senior Scientist: Why should you appoint a DS?

However you do it, appointing or hiring, full-time or part-time, a DS takes people and financial resources. Significant resources. It has been estimated that DM for a major cruise takes up to six months of staff time to do properly. We believe that the investment is worth it, to assist you as project leader and save your time, to ensure that the data set is well documented, to publicize your experiment, to gain access to other comparable data sets. Funding agents are well aware of the need for well documented data and metadata so they should be supportive. I repeat, from personal experience, having a DS assist you will save you significant amounts of time, which I find the most compelling reason, as well as pats on the back for the quality of your data.

For the Young Scientist: Why should you agree to act as DS?

Just finished a PhD? Desperate to write papers? Why waste time helping others with data management? My answer is that you should find talking to other PIs on the cruise/project interesting, and you will learn a lot if you are to understand in some detail what they do. That will broaden your outlook and may well give you ideas for your own research. It will give you management experience (persuading PIs to produce metadata, assisting them without dictating, helping them to deliver, keeping them to deadlines). All these will look good on your CV, help you at interview and make you more employable. You might even get paid!

Much of this cookbook is written as though you are the DS. Please read it even if you aren't! It will help you to understand what is needed. Quality control starts with the PI and is ultimately his/her responsibility.

Stages of Data Management

Project Planning: early stages

During initial Project write-up and budgeting

Ideally, 5% to 10% of a research project grant should be earmarked for data management (e.g. Glover *et al.* 2006). As a rule of thumb, data from a single multidisciplinary biogeochemical cruise will require between three to six months of post-cruise data management effort depending on the complexity and novelty of data types and the number of projects and PIs involved in that cruise.

What about collaboration with other national projects/initiatives? Can the data management effort be shared between multiple projects? For example, contact other PIs and consider pooling resources to fund a data management post, especially if a cruise is shared by multiple projects or teams from multiple institutes.

Contact a national/local data centre, provide them with quantitative and qualitative details of planned fieldwork and data types to be archived and compiled; ask them for an estimate of cost in terms of manpower and archive space. If no local contact and/or no local/national data centre is available, enquire from well established data centres in your main field of research.

Estimate DM needs by gathering initial project information

An estimate of the DM needs for the project can be started by collecting as much information about the proposed work as possible. For example, get listings of planned ship and cruise dates, get a list of all participants and proposed field work, and create a listing of the general types of data expected (paying special attention to the physical data storage needs). You can use the metadata templates in Appendix C to help with the initial gathering and tracking of the information at the project, fieldwork, and dataset levels.

Develop Rules of Information Exchange and Interaction

What are the rules for data exchange between participants within the project? What are the rules (and time scales for release) for data exchange and release with outsiders? Seek mutual agreement with participants, give them a say, allay concerns, and show benefits.

Project Planning: late stages

Establish a Web Identity for the DM component

A project Web site acts as a focal point for project participants and people looking for results and data from the project. Be sure this Web site also includes data management and data policies for the project. For example:

- The project Web site should include an informative

data management section, for example see the AMT [3] and BIOSOPE [4] projects.

- Prepare a DIF (Directory Interchange Format) record for the project based on information gathered in the Project Metadata template (see Appendix C: Project Metadata template). This DIF should be submitted to IMBER'S GCMD Portal, see also guidelines in Appendix B) to increase its visibility to the internet and/or searching software.
- Be sure to clearly publish the project data policy and data release time scales, for the sake of data user and provider.
- Be sure to provide information about planned fieldwork, creating dynamic fieldwork and dataset inventories.
- Provide links to other relevant online data resources.

Initial cut at Data and Metadata Handling

Get started on planning the management of the metadata associated with the data collected during the project. An excellent starting guide is available at the Marine Metadata Interoperability (MMI) Web site. [5]

The DS should thoroughly assess and review the full range and types of data anticipated from the project. Where they will be stored during the cruise and after the cruise, what will be needed to ensure data backup, what are the storage size (kilobytes or terabytes?) and data formats? Is the current project DM manpower and computer-power adequate?

Cruise Planning: before the cruise

Interact with all PIs on the cruise

Your most crucial role as DS is to talk to all PIs ahead of a cruise. Sell to them the importance of record keeping. Convince them that you will help them, not just add to their work load in form filling! Help them develop forms. Discuss with the cruise/project Principal Scientist (PS) how to ensure unique station and bottle references. Agree with the PS which records he/she would like you to keep. Assess data quantities, how each PI plans to store data, security, etc. Ideally this should all be done ahead of the cruise. In practice, it is hard to catch people on land, between meetings, at different places of work. Expect to have another go once you are at sea.

Pre-cruise Checklist

- Request the Principal Scientist (PS) to organize a cruise or fieldwork planning meeting with data management as one of the agenda items.
- Prior to this planning meeting, help all of the cruise participants provide a list of measurements to be made, along with methodology, sampling and analytical instrumentation; quantity of samples or volume and frequency of data; what analyses will be made on board, what will be made back in the laboratory; time scales for availability of the primary data after the end of fieldwork.
- Give a presentation on data management during the meeting. This is an opportunity for DS and/or PS to explain the rules and responsibilities of each participant and emphasize that data management should be an integral part of the project (not an after-thought). Show your draft of metadata for the cruise (see example in Appendix C: Fieldwork Metadata template) and your draft template for data set metadata (see example in Appendix C, Dataset Metadata template). Be sure to ask the audience/cruise participants for feedback and input.
- Present the data management setup for the cruise; stress the importance of (a) data archiving, (b) unique sample identifiers to be used by all, (c) accurate, precise and diligent space/time referencing, (d) central log keeping for gear deployments and underway discrete sampling.
- Interact with all PIs: learn as much as you can about the way they will operate; help them develop log sheets for their own use ; ensure that the key metadata information is correctly logged even if not required by the scientist for his/her own work.

Cruise Planning: during the cruise

This section is relevant to everybody, even though it is addressed to the DS. Note for example the importance of keeping good records, logging times accurately from the ship's clock, taking copies of your records and backups of your data. Yes, I know these are all obvious, but do you follow best practice all the time? I bet you don't.

3 AMT Project: http://www.bodc.ac.uk/projects/uk/amt/

4 BIOSOPE Project: http://www.obs-vlfr.fr/proof/vt/op/ec/biosope/bio.htm

5 MMI Web site: http://marinemetadata.org/guides/mdataintro/gettingstarted

A cruise is an invaluable opportunity to discover and document what PIs are doing. Most researchers are pleased to talk about their work. As DS, this is to your own advantage, as it broadens your knowledge, and it is to each PI's advantage, as you can help to ensure the quality and integrity of their data sets. So the most important role of the DS is to talk to all participants and document what they tell you. This includes the Principal Scientist/Project Leader (PS), of course. Other key roles are to keep track of central records, ensure that the primary data are accurate and complete, ensure all data are backed up and help with data exchange problems.

Interact with all other members of the cruise

Talk to all personnel. You should have done this before, but there is more focus once you are at sea and it is easier to catch people! Early in the cruise, try to learn what they do, how they keep records, how they store data. Update your drafts of DATASET metadata for each PI. How is their apparatus calibrated? How reliable are their standards? Are their data uniquely referenced to master variables (time, latitude, longitude, depth)? Is there any chance that bottle number (the number stamped on a sample bottle) could be confused with the position of the bottle on the rosette? If you see room for confusion, suggest diplomatically how their record keeping or standardization could be improved. But of course tread carefully. A new PI or grad student may be grateful for your suggestions. An experienced PI may not welcome any suggestion that he could do better! Ask first why he/she does things a certain way.

Back-up PI records (metadata and data)

All records are at their most vulnerable before they are copied for the first time. If a PI is nervous about allowing you to copy his/her unique, personal, uncorrected notes, you will need to reassure them that (data) security is your only objective and that the information is secure with you, and that you will not copy it to anybody else unless the PI authorizes you to do so.

As diplomatically as possible, take copies of each PI's records (keep your USB stick handy - or photocopy their notes and forms where necessary). Ensure that all data are backed up regularly - thus take copies of (e.g.) spreadsheets held on personal computers or laptops (I repeat: giving firm assurances that this is strictly to aid their security, you won't copy to others, etc).

Improving accuracy is your other card. If you check the PI's records, you can query errors, wrong station numbers, unlikely depths or dates, etc, thus making corrections while they are still fresh and easy to catch. Sharing data leads to better data.

Data Security

Researchers may be protective of their own records, and dubious about letting you copy information. To allay their concerns, you need to emphasize that you will not pass on any of their records without their permission, thus keeping their data secure. Point out that data are most vulnerable when first collected, and that is a major reason why you wish to copy them. Until you have photocopied the scrap of paper on which they have noted which of their bottles leaked during sampling, that scrap of paper could easily be blown overboard. (Don't laugh - it can happen all too easily.)

It is also notoriously easy for records held on a personal computer to be lost. They can be erased by mistake or corrupted by a disc crash. Here is a true conversation, after a cruise:

> "Please may I copy your spreadsheet?"
>
> "Yes, it is on my PC."
>
> "Can I copy it now?"
>
> "Er no, my student has taken the PC on a field trip to France. And, er, the student has just emailed me that the hard disc has been corrupted."

Moral is, of course, back up the computer regularly and copy vital files BEFORE the end of the cruise.

Keep good records for the cruise

Also keep track of central records, such as the Bridge Log, Station List, list of personnel and their responsibilities, and primary (shared) data such as CTD rosette firing information, and who drew what water from which bottles.

Flesh out your own list of PIs, data types, quantities of data, methods of data collection, calibration methods, all metadata needed to ensure data are fully described. Keep your own spreadsheet up to date on what you have obtained from PIs, and what remains to obtain. Make your own notes on each data type.

- **Maintain (and display) an event log:**

  A sure way to gain the confidence of researchers is to show that you are there to help them. If, during the

cruise, they can see that you are keeping and posting records that help them, then this is a good plus for you. For example, it is awfully easy to come on watch at 0400, half asleep, and write down the station number wrongly. Put up a central list, on which the watch leader enters all events - type of measurement, start and end times, station number, plenty of notes about problems, delays, etc (e.g. 0503 - CTD inboard, end of station 123; 0515 bongo net deployed). So you need to create a template with columns for needed information, print off lots of blanks, and ensure the watch keepers know where they are so they can renew each sheet when full. This event log, of times and facts, is different from the cruise narrative (diary) which the PS will probably write (aided by the event log). You can find similar discussion and an example of an event log in the BCO-DMO document. [6]

On your watch (i.e. regularly, once or twice a day), type the hand-written event log data into a spreadsheet and check it. Does it agree with the Bridge Log? Can you enter events that the watch leader has missed but the Bridge has recorded (e.g. down time for engine problem or weather)? Print out and pin up completed pages so that a PI, often catching up a couple of days late after fixing some urgent malfunction, can update his own records from your master list, check he has the station numbers and times right, etc. Putting up your printed version also allows others to verify your numbers and correct errors and misunderstandings. There will be plenty.

- **Ensure "Primary Data" are quality-checked and readily available:**

  Primary data (e.g. navigation data, station data, etc.) are needed by everybody. Quality-control these yourself or ensure somebody else is doing this thoroughly. Use plots to check for errors (jumps in station position, backward time steps). Ensure all scientists use the master (quality-controlled) data - for example, many scientists will write down the lat and lon and time roughly for a station for their own reference, but these values must be replaced by the proper values sooner rather than later.

  - Time: Time is the master reference variable. In these days of GPS, the navigation file will contain highly accurate time, latitude and longitude, so that positions can be accurately obtained by looking up lat and lon given an accurate time. Hopefully, any research ship will have an accurate clock with slave clocks in each laboratory showing time in an agreed time zone (often GMT). Use that clock to check and write down all times. Do not use personal watches. They could be wrong and may well be set to the wrong time zone, for example if the ship is working in a different time zone from that being used for scientific logging. It may be true that for many purposes time does not need to be all that accurate. A minute or two may not matter for the time a net was hauled or a CTD reached maximum depth. But for some purposes fractions of a second matter, such as obtaining accurate currents from an ADCP using ship's position and heading data. So make a habit of recording time accurately, at least to the nearest minute, more accurately if necessary.

  - Positions: It is tempting to write down lat and lon from a lab display, and log sheets often have a space to enter position. Lots of errors arise this way. Written lat and lon are indeed useful, but time is the master variable and the one to double check. A useful task for the DS when typing up the event log is to add accurate positions to major times (e.g. CTD at maximum depth) by extracting them from the navigation file. Perhaps this can be automated. Another check is to plot a map of station positions. It is a safe bet that one or two station positions will look odd because of a transcription error. Overlay station positions on a track plot to double check. It really is good to carry out these checks as soon as possible. Otherwise errors will propagate as people copy the wrong values, transfer them to a data centre, etc. You can be sure that someone will map station positions eventually, probably Fig. 1 of your paper, and then you'll curse when an error is obvious!

  - Station Positions: Ships often drift during a supposedly fixed location station, like a net or CTD cast. For some purposes, it may be necessary to know start, middle and end positions (easy if times are logged) but most often the convention (for which there are good reasons) is to use the time and position when the instrument reaches its maximum depth, the "bottom" time.

  - CTD bottle file: On many cruises CTDs are a staple measurement, typically associated with a 24-bottle rosette sampler. While the physicist is primarily interested in the CTD (pressure,

6 BCO-DMO document: http://bcodmo.org/files/bcodmo/BCO-DMO_best_prac_v1d2.pdf

temperature, salinity), biologists and chemists depend on bottle samples for many measurements (chlorophyll, Fv/Fm, 234Th, dissolved iron, nutrients, oxygen, species composition, etc, etc) as well as wishing to know the CTD parameters (P, T, S, possibly also oxygen, light,...). An important duty of the PS is to ensure that sampling protocols are agreed by all participants at the start of the cruise and watch leaders must ensure that protocols are rigorously followed. In what order must samples be drawn to avoid contamination? Who wants to sample which bottles and at what depths? Bottle labeling has pitfalls too. The position on the rosette determines at what depth the bottle was fired, and the technician controlling bottle firing must write down carefully times and depths of bottle fired, particularly noting any misfires. But the label on the bottle is also important in case a bottle leaks or appears contaminated. If a bottle is moved or replaced the details must be carefully documented. While bottle sampling is taking place, one person should be delegated to watch and check that the samplers are working in the right order and drawing from the correct bottles. You can be sure that problems will arise later, however much care you take: "that bottle can't have been sampled at 5000m, the temperature is 15°C". With a detailed paper trail, hopefully problems like this can be resolved. The next problem is that it is non-trivial to create a single file (could well be a spreadsheet, because only 24-36 data cycles per CTD cast) with correct values for all the sampled parameters. This is because the samples for each parameter are separately drawn and analyzed. Some samples (e.g. for oxygens, nutrients) need to be analyzed within hours for high-quality data, others (e.g. elements with long half-lives) cannot be counted for months. Yet other samples (e.g. chlorophyll) may be reworked several times with improved standards, each time improving accuracy. Someone, likely the DS, needs to maintain the CTD bottle spreadsheet to ensure that it is complete and up to date. The CTD software should make it straight forward to transfer CTD parameters when the bottle was fired into the bottle file. Participants will have to be chased/cajoled/bullied to obtain other parameters as soon as possible. Excuses will be myriad: 'the constant temperature lab is at the wrong temperature; the autoanalyzer is acting up again; I am waiting for a full load for the sampler.'

Assist the Principal Scientist

Establishing yourself as a contact path between the PS, PIs, data, and primary data makes all of your lives easier. In particular, it is important to agree with the PS which tasks you should be responsible for from the list above. The PS may want to do some tasks himself (though both PS and DS need to talk to PIs). There may be more tasks than you can do yourself given other commitments, so the PS or DS may have to delegate others to help.

- **Assist with Cruise Summary Report**

  Working with the PS to prepare the Cruise Summary Report (CSR) is an opportunity to co-review all events and records of the cruise, ensuring that all PI data and events are accounted for and verifying you both have identical records of the cruise and events. The CSR is discussed in Appendix A and much more detail is available at ICES ROSCOP Web site.[7] Much of the detail requested by the CSR you will have already in your PLATFORM and DATASET metadata forms. Assembling the cruise report is significantly easier if the materials and information are gathered together during the cruise, and definitely before the ship docks. The DS should work with the PI to ensure that these elements have been provided by the cruise participants before they leave the ship.

- **Time scales for data calibration, analysis and delivery**

  In preparation for post-cruise tasks, discuss (and document) realistic time scales for data delivery with each PI. Alert PS if you think PI is being over-optimistic, or is in need of assistance. Remember, individuals always underestimate how long it will take to deal with their data, and are quickly overtaken by other responsibilities as soon as they return to their offices.

After the Cruise: early stages

One of the first post-cruise responsibilities is to submit the cruise summary report to a national or international data center. Web access to the Event Log for your co-participants and national data center would also be useful.

Primary Data

Make the quality controlled primary data available to all participants as soon as possible, or they will use their own versions. The navigation file (time, lat, lon) is one of these, the station list is another.

7 ICES ROSCOP Web site: http://www.ices.dk/Ocean/roscop/index.asp

However hard you try, there will probably be gaps and errors in the CTD bottle file at the end of the cruise. Parameter values for the last few stations are not worked up yet. Counting is not complete. The DS needs to note what values are outstanding or not final and when they can be completed. Import new parameters as analysis is completed (this may be months after the cruise for some parameters). Follow up until the spreadsheet IS final.

Completion of the Cruise report

The detailed narrative cruise report was hopefully started during the cruise, and should be completed as soon as possible after the cruise. The DS should assist the cruise Principal Scientist in this task. For guidelines on compiling informative cruise report see the BODC Cruise Compilation Guide [8] and also Appendix 1.2 of the BCO-DMO Best Practices document. [9]

Archive individual master files

Work with the project participants to generate a complete and quality-controlled master file or master set of files of their data as soon as possible after the cruise or after the analyses of the samples have been completed. It is important that the master file(s) be easily readable by various data handling tools or data management systems and unequivocally understandable by other people, including non-specialists.

We recommend scientists to follow the best practices compiled by Cook *et al.* (2001) and revised by Hook *et al* (2010) for environmental data sets.[10] They identify seven best practices all of which are relevant to the marine community and to IMBER:

1. Assign descriptive and unique file names
2. Use consistent and stable file formats for tabular and image data
3. Define the contents of your data files
4. Use consistent data organization
5. Perform basic quality assurance
6. Assign descriptive data set titles
7. Provide documentation

Using the guidelines and resources listed on the MMI Web site [11], the DS should help the scientist identify the most relevant standards and vocabulary to label and document their data.

Once the masters have been generated, the write permission on the file should be removed and the files and their associated documentation should be saved in an archive. Copy of the archive should be made and saved at multiple locations and/or on different media to guarantee future access to the data (personal computer, local network, online archive, data centre, or external storage media). Ideally one copy should be sent immediately to a well-established data centre who could then apply its own data tracking and safekeeping procedures.

Keep track of progress

Interact regularly with each PI to help and advise them, as well as to keep yourself up to date with their progress. Take copies of their data sets each time these are significantly changed (edited, calibrated, validated). Ensure successive versions of data are clearly identified.

Data validation and specialist data centres

Data validation involves comparison of the data from your project with other similar data. For example, some measurements might be offset because of an error in procedure, standardization or similar. For some data types, validation can be done by specialist data centres. In return, they are pleased to have copies of new data to enhance their global data bases.

After the Cruise: late stages

Your work is almost done, but first you must ensure that all data from the cruises (and project?) are submitted to long-term archival data centres. Once the data have been submitted, check that those data have been fully processed/archived by the data centres, perhaps by trying to search for and acquire a copy of the data via the data centre (e.g., using the method a typical user would use to access the data either through direct request or through an online portal). Make sure you can find all of your submitted data, then review the metadata (i.e., project title, methods, author accreditation) and the data records to ensure that all files, data, and metadata are presented.

8 BODC Cruise Compilation Guide: http://www.bodc.ac.uk/data/information_and_inventories/cruise_inventory/documents/bodc_guide_cr_compilation.pdf

9 BCO-DMO Best Practices document: http://bcodmo.org/files/bcodmo/BCO-DMO_best_prac_v1d2.pdf

10 Best Practices document: http://daac.ornl.gov/PI/bestprac.html

11 MMI Web site: http://marinemetadata.org/guides

ACRONYMS

CSR	Cruise Summary Report
DIF	Directory Interchange Format
DM	Data Management
DS	Data integration Scientist
GCMD	NASA's Global Change Master Directory [12]
MMI	Marine Metadata Interoperability Project [13]
PI	Principal Investigator
PS	Principal Scientist (e.g. scientific cruise leader)

REFERENCES

Cook, Robert B, Richard J. Olson, Paul Kanciruk, and Leslie A. Hook. 2001. Best Practices for Preparing Ecological Data Sets to Share and Archive. Bulletin of the Ecological Society of America, Vol. 82, No. 2, April 2001.

Glover, D.M., C.L. Chandler, S.C. Doney, K.O. Buessler, G. Heimerdinger, J.K.B. Bishop and G.R. Flierl 2006. The US JGOFS data management experience. Deep-Sea Research II 53: 793-802. [14]

Hook, Les A., Suresh K. Santhana Vannan, Tammy W. Beaty, Robert B. Cook, and Bruce E. Wilson. 2010. Best Practices for Preparing Environmental Data Sets to Share and Archive. Available online (http://daac.ornl.gov/PI/BestPractices-2010.pdf) from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, U.S.A. doi:10.3334/ORNLDAAC/BestPractices-2010

12 GCMD: http://gcmd.nasa.gov/

13 MMI: http://marinemetadata.org/

14 Glover *et al.* 2006: http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6VGC-4K1G54V-3&_user=3615566&_rdoc=1&_fmt=&_orig=search&_sort=d&view=c&_version=1&_urlVersion=0&_userid=3615566&md5=d5d93aaf9a6120ef324a480c19e363e3

Appendix A: The Cruise Summary Report

The Cruise Summary Report (CSR) previously known as ROSCOP (Report of Observations/Samples Collected by Oceanographic Programmes) is an established international standard designed to gather information about oceanographic data collected at sea. The form was designed in the late 1960s by the Intergovernmental Oceanographic Commission (IOC) of UNESCO and redesigned in the late 1990s. It has been widely adopted by many IOC member states and related organisations such as International Council for the Exploration of the Seas (ICES/CIEM) and, more recently, the Partnership for Observation of the Global Oceans (POGO).

In many countries with a national oceanographic fleet, Principal Scientific Officers (PSOs)/Chief Scientists are asked to submit a CSR to their designated Oceanographic Data Centre. There, the information is checked and entered in a national database before being submitted to the international CSR database maintained by ICES. In the absence of a national collating centre, the PSO can submit the information directly to ICES. [15]

Cruise Summary Reports will enable IMBER to maintain a low-level inventory of data collected during research cruises in the frame of the projects and activities it has endorsed. It is therefore requested that PSOs notify the IMBER International Project Office (IPO) has soon as a CSR form is submitted and send a duplicate copy of the completed form to the IMBER IPO.

In countries where a different type of reporting form is used, a copy of the form should be sent to the IMBER IPO where advice will be sought on the best way to convert one form into another and avoid duplication of effort. Alternatively, the information may be copied and pasted into a CSR form before being submitted to a collating database and to the IMBER IPO.

Cruise summary reports should be submitted no more than two weeks after the end of the cruise. They DO NOT replace the conventional more detailed Cruise Report which should be completed and submitted to a national or international repository no later than 6 months after the cruise.

A template example of the CSR form as a WORD document can be downloaded from the ICES Web site [16] or from the web sites of many National Oceanographic Data Centres.

Online resources related to Cruise Summary Reports:

- CSR form (+ parameters codes + Marsden square)
- CSR form (smaller file than above) [17]
- CSR parameter codes [18]
- List of Marsden squares for cruise location [19]
- International Hydrographic Bureau Sea Areas (IHBSA) [20]

15 ICES Cruise Summary Report information: http://www.ices.dk/Ocean/roscop/index.asp

16 ICES CSR form: http://www.ices.dk/Ocean/roscop/ros-doc.doc

17 CSR short form: http://www.bodc.ac.uk/data/information_and_inventories/cruise_inventory/documents/new_csr_form.doc

18 CSR parameter codes: http://www.ices.dk/Ocean/roscop/par-cod.htm

19 Marsden Squares list: http://www.ices.dk/Ocean/roscop/msq.csv

20 IHBSA: http://www.ices.dk/ocean/codes/ihb.htm

Appendix B: How to create DIFs for IMBER

The IMBER Data Management Committee recommended that IMBER adopt the Directory Interchange Format (DIF) as a discovery metadata standard. The big advantage in using DIF is that records can be easily created and managed through NASA's Global Change Master Directory (GCMD). A customized metadata portal within GCMD has been set up for IMBER and can be accessed at the GCMD IMBER portal. [21]

The aim of the portal is to provide a central access point for searching information about the data collected in the frame of IMBER's projects including regional and national activities, joint activities with other programmes and of course, endorsed projects. The creation of DIF records is the responsibility of the project leaders, project data scientists or project participants; however the IMBER Data Liaison Officer at the International Project Office will provide assistance and be responsible for overseeing the quality and consistency of the information provided.

The purpose of the present guidelines is to inform project leaders, data scientists and other IMBER participants on IMBER's minimum requirements. In order to make the IMBER portal useful as a tool to discover information about IMBER data it is important that the information be reported accurately, consistently and comprehensively. General information about writing DIF records can be found in the GCMD DIF Users Guide. [22]

There are two ways of contributing to IMBER metadata portal:

- linking an existing DIF to IMBER
- creating a new DIF and linking it to IMBER. [23]

In both cases, from the IMBER portal click on "Add to IMBER portal" [18] which transfers you to the GCMD DocBuilder site. Either type in a name for a new DIF or select "work with an existing document" then type in the name of the existing document. The button then takes you into DocBuilder where you can work on your DIF, which is automatically saved. Note which fields are essential and fill in step by step, using the guides mentioned above. Start by clicking on Project and selecting IMBER. The drop down menu Document/Preview document display is very useful to see the final layout once you have entered information (includes a map when you have entered geographical information (Spatial coverage). Use Document/Validate to check your fields, but note that new key words have to be checked and approved by GCMD once submitted. Once complete, submit the document to GCMD, who will forward it to the IMBER IPO for checking before it is made public.

In the frame of IMBER, three levels have been identified at which a DIF record could be created: the project level, the campaign/cruise level and the individual scientist level. Below are details about creating DIFs at these three levels.

Metadata DIF record at Project level

In addition to following the guidelines providing by GCMD [24], DIF authors should be aware of the following:

- At the minimum, each project/program related to or endorsed by IMBER should have a DIF record in GCMD with a link to IMBER in the "project" field.
- If a DIF has already been created in the frame of another programme (for example IPY, GLOBEC, GODAE, SOLAS, national activities) and needs to be made visible to the IMBER portal then, all you need to do is to add IMBER to the "project" field.
- To link a DIF to IMBER, select IMBER from the drop-down list when creating or editing the DIF. DIF records can be linked to more than one project.
- DIF records should clearly indicate the data management and data curation arrangements for the project.

Metadata DIF record at Cruise or Data Collection Activity level

IMBER is currently investigating means by which CSRs can be easily converted to DIF therefore avoiding duplication of effort and enabling information about cruises to be searchable via the metadata portal. For activities other than cruises, DIF records must be created

21 GCMD IMBER portal: http://gcmd.nasa.gov/portals/imber/

22 GCMD DIF Users Guide: http://gcmd.gsfc.nasa.gov/User/difguide/DIF_Guide_2010.pdf

23 User accounts are now required for adding and modifying descriptions in the GCMD. Please contact the GCMD User Support Office (gsfc-gcmduso@mail.nasa.gov) to request a new account, or visit the following page: https://users.eosdis.nasa.gov/urs/welcome.do

24 GCMD DIF Tips: http://gcmd.nasa.gov/User/difguide/dif_tips.html

for each activity and linked to IMBER and to the parent project if applicable.

Note that in order to link a record to a project or a campaign, you will need to add the campaign or project to the list of GCMD projects (this is independent from creating a DIF record to describe the project or campaign).

Metadata DIF record at the Individual Scientist level

Scientists may want in some instances to create a DIF record for their individual data sets. If this is done, we recommend either to create these individual records as child records of the project or of the field campaign, whichever is the most appropriate or, if the record does not easily fit in a parent-child relationship, then to ensure it is linked to IMBER in the project field.

Online resources related to DIF records, GCMD and discovery metadata:

- Usage versus Discovery metadata [25]
- DIF writer's guide [26]
- DocBUILDER authoring tool (you must be logged in first) [27]
- Create an IMBER DIF record (you must be logged in first) [28]

25 Usage vs. Discovery Metadata: http://marinemetadata.org/guides/vocabs/voctypes/cvusagevsdisc

26 DIF Writer's Guide: http://gcmd.nasa.gov/User/difguide/difman.html

27 GCMD Doc Builder: http://gcmd.nasa.gov/DocumentBuilder/Home.do?RequestAction=Help

28 GCMD IMBER DIF: http://gcmd.nasa.gov/DocumentBuilder/Home.do?Portal=imber&MetadataType=0

Appendix C: Metadata Templates

Project Metadata Template

**The templates presented in this section were modified, with permission, from templates originally developed by BCO-DMO.*

PROJECT description metadata for IMBER

[Enter as much information as possible; project name and description are required.]

Project Name:

- **Acronym:**
- **Program Name (if relevant):**
  [larger program (other than IMBER) with which this project is directly affiliated; Ex. BASIN or OCEANS 2025]
- **Affiliated program(s):**
  [list any additional national or international programs with which this project is affiliated]
- **Project url: http://**
- **Funding:**
  [agency and award number; e.g. NSF-OCE 9999999]
- **Lead PI name and contact information:**
  [full name and current email, mailing address, phone, etc. or name of the head of the Steering Committee and contact information, URL, etc.]
- **Co-PI name(s) and contact information:**
  [repeat names and contact information as needed]
- **Contact name:**
  [best person to contact with any questions about this project]
- **Start date:**
  [of field work]
- **End date:**
  [of all field work]
- **Logo url:**
  [or attach image file]
- **Geolocation:**
  [general description of study area; Ex: Sub-Antarctic waters 48 S 173 E.]

Project Description:

- **Detailed description (or Science Plan document)**
- **Related files:**
  [include supporting documents: science plan, original proposal, background publications]

Fieldwork/Cruise/Activity Metadata Template

**The templates presented in this section were modified, with permission, from templates originally developed by BCO-DMO.*

FIELDWORK/ACTIVITY description metadata for IMBER

[Enter as much information as possible, platform name and type are required, deployment identification required for platform type=vessel (e.g. cruises).]

Fieldwork/Activity Name:

Fieldwork, Activity, or platform type:

[vessel, mooring, satellite, aircraft, ROV, towed vehicle, submarine-manned, submarine-unmanned, island, model, synthesis, mesocosm]

Deployment identifier:

[cruise ID, mooring ID, dive number]

- **Synonyms:**
  [other names used to refer to this cruise or deployment]
- **Coordinated Platforms:**
  [platform deployments coordinated with this one]
- **Project:**
  [associated project]
- **Principal scientist name and contact information:**
  [full name and current email, mailing address, phone, etc.]
- **Co-Principal scientist name and contact information:**
- **Contact name and contact information:**
  [best person to contact with any questions about this activity]
- **Start date:**
  [e.g. date ship left port; date mooring deployed]

- **End date:**
  [e.g. date ship returned to port; date mooring recovered]
- **Location:**
  [general description of study area; Ex: Sub-Antarctic waters 48 S 173 E.; lat/lon bounding box]
- **Related files:**
  [include supporting documents: science or cruise plan, Cruise Summary Report, activity report, shiptrack maps]

Dataset Metadata Template

**The templates presented in this section were modified, with permission, from templates originally developed by BCO-DMO.*

DATASET description metadata for IMBER

[Enter as much information as possible; PI name, cruise ID and dataset description required]

Dataset Short Name:

[Preferred short name (20 characters or less) for the dataset]

Dataset title:

[Brief sentence describing these data; preferably; less than 60 characters]

Dataset description:

[Brief summary description of the dataset, its context and content]

Deployment/Activity identifier:

[Cruise ID, mooring ID, dive number]

- **Deployment/Activity Synonyms:**
  [Other names commonly used to refer to this cruise, deployment, activity]
- **Project:**
  [The name of the project with which these data are directly associated]
- **Funding:**
  [Agency and award number; e.g. NSF-OCE 9999999]

Originating PI name and contact information:

[Full name and current email, mailing address, phone, etc. for the PI associated with these data]

- **Co-PI name(s) and contact information:**
- **Contact name and contact information:**
  [Best person to contact with any questions about these data (could be PI, post-doc or assistant]
- **Location:**
  [General description of study area; ex: Sub-Antarctic waters 48 S 173 E.; lat/lon bounding box]
- **Parameter names, definitions and units:**
  [If these are not explained in the data files, please include definitions and units here]
- **Sampling and Analytical Methodology:**
  [Include written description of methods or separate files with description of sampling and analytical methodology; please include name and description of sampling and analytical equipment and instrumentation and details of quality assurance and control procedures]
- **PI Notes:**
  [Can include anecdotal comments, notes or details regarding data quality]
- **Related files and references:**
  [Include any useful supporting documents; e.g. separate files or published papers with description of sampling and analytical methodology]

IMBER

The Integrated Marine Biogeochemistry and Ecosystem Research project is a multidisciplinary project of the International Geosphere-Biosphere Programme (IGBP) and the Scientific Committee on Oceanic Research (SCOR). Both IGBP and SCOR are interdisciplinary bodies of the International Council for Science (ICSU).