

Appendix I Session Reports

Session on Preceding Work (Rapporteur: Katy Hill)

Roy Lowry introduced the session and background. There wasn't a strong attendance at the Data Management session in Banff, and the meeting was dominated by data managers. There is a need to make sure we are on the right track, so this session will be used to make sure that the Banff report reflects the thoughts of the community.

What recommendations should be taken forward from the Banff Report?

- Establishment of a Data and information management unit at the outset.

This is something that everyone seemed to agree with.

- Development on distributed scalable data management.

You don't want to depend on getting all data to one place.

Jim Swift: Why distributed? I can see how it works, but if you have a supercentre then that is fine.

Roy Lowry: Pragmatic solution based on funding

Jim Swift: What is exactly the key point? They have worked, but are they essential? Some form of Data Information Unit is essential.

- Adoption of standards to facilitate interoperability of data and information.

Roy Lowry: JGOFS didn't integrate more because BODC didn't have the resources to rebuild interfaces. Is now the time to standardise measurements and formats?

Jim Swift: Measurements change and evolve through the timeframe of a program

Ferris Webster: Careful about standards because you don't want to limit scientific ingenuity.

Standard nomenclature is fine, but you don't want to straightjacket project scientists.

Roy Lowry: Standard metadata vocabulary. If you produce measurements at a higher level of accuracy, you are working to a dictionary, you add another line to the dictionary.

Jim Swift: Allow for evolution, and yet promote interoperability.

Roy Lowry: Standards have to be organic, developing things: They shouldn't be a straightjacket. Some areas have more developed standards than others, but it is something that should be promoted across the board.

- Utilisation of existing infrastructure but with additional resources to ensure it fulfils international rather than national standards and specifications.

With JGOFS there was funding available for the national effort, but this did not extend to developing international interoperability.

- Provision of services and data access that match the needs of scientists

Wendy Broadgate: What about different ways in which you can encourage scientists to submit data. Can we add the possibility of giving scientists incentives for submitting data to this point?

Roy Lowry: Yes definitely.

Bernard Avril: add scientists and other end users.

Roy Lowry: good point as this could extend to local/national government etc.

- Provision of data through alternative media e.g., CD Rom for those without internet access.

Jim Swift: There is no point in specifying media, because it evolves.

Roy Lowry: Change to "a leading edge technology and a universally available technology"

- Development of a close working relationship between data managers and scientists through means such as 'end to end' project data management and the provision of data access tools.

Jim Swift: Data management ends long after project ends.

RECOMMENTATIONS:

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

- Data Sharing within the programme, between programmes and the entry of data into the public domain.

Roy Lowry: Aware of data policies which state that all data will be there within two years. You need a get-out clause here. These are mechanisms as opposed to rules. e.g., Tracer data – don't get their first numbers for ~18 months.

Jim Swift: Knowing more information about data is the key. Then, at least people know it is there and can work on getting it released.

Data quality issues

Jim Swift: Data CONTENT and quality issues. Programs need to first address what their data are.

Bernard Avril: isn't that more the core science of the program?

Roy Lowry: this is part of the data policy. But, it is definitely a science issue.

Casey Ryan: It is a very daunting prospect for new projects such as SOLAS. It is not possible to know what kind of data will be collected in the next 10 years.

Roy Lowry: specifying minimum data content. WOCE put a cap on this, JGOFS didn't.

Long-term security of data

Ed Urban: Is this limited to WDCs?

Ferris Webster: There are other non-WDC data centres with long-term stability. For example, the PO.DAAC does not submit to the Oceans WDC. The sheer volume of satellite data held at PO.DAAC makes this prohibitive.

Roy Lowry: you need to make sure that they have the resources to take the data onboard and have long-term security of storage.

2. All new programs should dedicate a resource to the development of a project metadatabase that will form the project data inventory. This should conform to appropriate international standards (e.g., ISO19115 for spatially referenced data) to facilitate integration and exchange information between programmes. Previous experience has shown that this resource is most effective if located in the IPO.

Jim Swift: In WOCE, the metadata resource was scaleable: We had the DIU, and then the DACs had their own metadatabases.

Roy Lowry: WOCE had a tightly constrained set of parameters. JGOFS was different.

Jim Swift: What is an inventory to one person, is a shell to someone else. Different people have different requirements from such a metadatabase.

Julie Hall: shouldn't that be part of the metadata anyway?

Roy Lowry: The IPO should be a portal for metadata. You really need a central portal somewhere.

Bernard Avril: Metadata levels are dictated by core science.

Roy Lowry: The reasons for linking metadata to IPOs is that information about the science is known best at the IPO.

Julie Hall: In IMBER, there will not be a set of core parameters; therefore, you can't constrain the boundaries for the metadata inventory.

Roy Lowry: OMEX an example. 4400 different parameters measured on one water sample. They used CSR (cruise summary report) forms.

Jim Swift: One person's signal can be another person's noise; the same goes for metadata.

How many metadata slaves do you have to hire?

Howard Cattle: Roy has a point for projects with a limited remit. CLIVAR and others go across discipline boundaries. The IPO act as an intermediary. The expertise is found within the panels.

Jim Swift: The IPO should ensure that a system is enforced appropriate to the needs of the project, and have the funds available to implement this.

Wendy Broadgate: The role of the IPO is to support the SSC in a practical way.

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

Jim Swift: Last phrase should be “data originators, data managers and data users”.

4. *National science programs should address data management in a credible manner including giving consideration to capacity building if appropriate.*

Jim Swift: How much should data management cost?

Roy Lowry: Figures tossed around: in EU – 5-10% of cost of project budget. But bureaucracy often caps this. 5% is generally a good number.

Jim Swift: Should we include this figure?

Stefan Rothe: It is dangerous to specify a figure.

Roy Lowry: Including allocation of appropriate resources and giving (?)

Bernard Avril: You can only recommend – you cannot enforce data management as you will exclude data sources.

Ferris Webster: remove national. Start... “Science programs should...”

Roy: National or project science programs. You need to make sure that data management is allocated at the level where the money is.

Wendy Broadgate: need to work on the wording of this to make it more explicit.

Katy Hill: Need to address data management at the level where money is distributed. Can this be written in?

Wendy Broadgate: Also at a labelling level.

Julie Hall: That is at a higher level: what is necessary to be endorsed?

Ed Urban: Capacity building is at the international level and not really addressed at the national level.

Holes and Missed Points?

Attention should be given for developing incentives for scientists to submit/share their data. This issue will likely be addressed later in the meeting.

Wendy Broadgate: Offering tools such as modelling, plotting (cartographic representation of data) etc.

Casey Ryan: Projects need to think carefully about what would provide effective incentives.

Liana McManus: It needs to be thought of as a two-way data exchange and need to address how to foster that exchange.

Roy Lowry: The two-way exchange is an incentive itself.

Wendy Broadgate: The right to ownership and first authorship of data is important to guarantee.

Roy Lowry: Needs to address Intellectual Property Rights. It is a data policy issue.

Session on the WOCE experience (Rapporteur: Bernard Avril)

Several needs were identified in relation to proper, credible end-to-end data management, after the WOCE experience:

- to clarify the human mechanisms to improve the communications between data management and data originators and users (e.g., encouragement of the participation of data managers on cruises; service-oriented attitude addressing the user needs; establishment of some support from a regionally responsible data manager for the countries with no adequate data management support)
- to promote the support, services, and benefits offered by the data management system
- to undertake a strong effort to motivate the PIs to participate to the data management effort when there is no national structure
- to clarify the range of activities in the international project. Multi-labelling of national projects is often the case.

- to change the attitude of some funding agencies which do not support data management or are not willing to contribute to international projects.
- to take into account the commonalities among projects. Each project could have its own data committee, and each can contribute to a partly common data management infrastructure (DACs, NODCs), not only within IGBP. The data management system needed depends on the science done... need to be driven by the science, not be a fixed template before the science is defined (not "one-size-fits-all" structure and mechanisms). In addition, WOCE DAC proposals were successful when they were tightly linked to WOCE science.
- to define the working mechanisms for the data management system and then define the data products that are expected. The data management system needs to be able to evolve
- to take into account the history of development of previous project data management
- to define, between projects, a common parameter dictionary
- to motivate PIs to participate in the data management system. What added value is given to them?
- to make the submission of data as simple as possible... and to keep asking until the information is obtained
- to promote project data management/end-to-end data management, maybe through a handbook of best practices in data and information management.

The "Data Committee" (with broad composition) should aim to ensure that the data management system addresses the needs of the scientists, nationally and for the international integration. Yet, the Data Committee should not include representatives from all participating countries, because small countries cannot support a data manager, and in some large projects the Data Committee would be too large to be efficient.

A basic question (from the agenda) was not fully examined: *Are data or peer-reviewed papers the project legacy?*

The reasons of the WOCE success were presented and discussed. They include:

- adoption of standards (e.g., netCDF, although the move from ASCII to netCDF was not easy),
- small set of parameters (physics and basic chemistry / biology),
- use of the DACs, with (mostly) only 12 data types

Yet, DAC-like bodies may not be appropriate for biological data. At least a minimum set of parameters (e.g., CTD, pigments, oxygen, nutrients) could be handled by DACs, but the DACs in WOCE were driven by the science. DACs could create quality-controlled data collections, across the cruises, and even the projects. DACs could be organized by data streams... DACs could also be a reference body for the isolated PIs without adequate national data support in their country. The data assembly and documentation is the hardest job. Quality control is easier when data are assembled and documented. Specialised DACs would be better recognized and used by PIs, helping them to better organize and use their own data and others' data. DACs should be set up with a clear set of services they can provide, and then some funding should be seek with some clear added-value science deliverables that could not be achieved without the DAC's work. In addition, the outputs from DACs should be disseminated as easily/rapidly as possible, because the scientists work with more than one data type.

On the other hand, JGOFS needed a unified parameter dictionary and a complete inventory of the data collection, and never managed to get them... because JGOFS was more complicated, and because no adequate project endorsement was applied. And the data managers needed to first understand the meaning of the parameters and then to prepare some standards. The information on the on-going activities was also a very difficult information to acquire, outside the DMTT-represented countries. It appears useful to think ahead before the cruise in order to homogenise the referencing of the activities (events) and sampling numbering among all cruise participants. This is much easier than to do it later with a cross-referencing table...

In conclusion, whatever the size and complexity of the project, there are always some mechanisms in data management that could be easy to set up and would be very helpful to all.

Session on Metadata Management (Rapporteur: Bernard Avril)

The “meaning” and value of metadata were examined: technical metadata (for automated system); semantic metadata (increase the human understanding of the data); discovery metadata (help to locate the data, DIF, MEDI, EDMED, FGCD...). Already, GCMD’s DIF can indeed handle a wide range of ecological parameters.

Metadata submission should be an important criteria in the endorsement of the PIs or country participation. It is recommended to uncouple the data management (which could be distributed) and the discovery metadata management (e.g., through WOCE DIU-type; centralized structure, helping to define the project as a clear entity, to increase the PIs recognition, and to preserve the data management system in case some data centres disappear; using the list of proposed data actually collected, and then cross-checking with data delivery). What, Where, When, Who, are basic information included in metadata;

The data management system should agree with some general principles and respond to the project-specific requirements for the data and metadata, and reciprocally display the benefit of the data management system (e.g., protection of the authoring, via a wide dissemination of the metadata and the protection of the actual datasets, until further notice by the PIs).

It is recommended that (raw) metadata be first handled by IPO until the data are actually available and then handled by an adequate N(O)DC, or a WDC.

It is recommended that reciprocity in the processes of data acquisition and data dissemination has to be a mainstream behaviour. It might require a cultural change.

Some items in the agenda were not fully examined:

- *Value of data without metadata?*
 - *Standardization for interoperability*
 - *Structure of a metadata repository*
-

Session on National Data Management Infrastructure (Rapporteur: Casey Ryan)

- End-to-end data management is a useful concept and should be reproduced and adapted. It is limited by scale.
 - Procedures have to be developed to operate this in countries without a data management infrastructure
-

Session on Data Policies (Rapporteur: Dawn Ashby)

Recommendations:

- SSCs are responsible for developing the rules for data sharing both within and between projects.
- As we are all ICSU bodies we should promote free and open access to data in line with ICSU policies.
- Existing data policies should be used to formulate a template data policy for future IGBP/SCOR projects

- The data management committee should report adherence to the data management policy to the SSC, who should consider those who do not comply on a case-by-case basis.

What existing data policy documents are available as models?

There are many existing data policy documents for IGBP programmes which can be used as a model, the ACSOE data policy was also raised as a good example which contains a 10-point summary detailing benefits to the scientists, intellectual property rights and submission time scales ranging from 6 months to 2 years, depending upon the discipline.

Who decides the rules for data sharing within a project?

It was agreed that this was the responsibility of the Scientific Steering Committee.

Who decides the rules for data sharing between projects?

There is an obvious need for sharing data between IGBP projects. The rules for sharing data before it enters the public domain should be determined by mutual negotiation of the SSCs of each project.

Is there a universally applicable template for a universal data sharing policy?

Yes, there are similarities between data policies that could be extracted to form a template policy.

What are the major differences in national data sharing cultures that need to be addressed by international projects?

Political and economic restrictions may restrict sharing of data from certain countries and this should be considered when developing data policies. Examples include India where the Ministry of Defence restricts exchange of certain hydrographic data from their EEZ, and Russia, where data in digital format are not allowed out of the country.

Are there additional issues to consider for non-marine data such as atmospheric or socio-economic data?

In the UK, the culture for sharing data was considered to be similar between atmospheric (BADDC) and oceanographic (BODC) data centres. Problems may arise with meteorological data, as it has commercial value. The Met Office provides data “at cost” but this is still too expensive.

There are technical difficulties in managing non-geospatially referenced and socio-economic data. There is a human dimensions WDC which was thought to contain very little data and it is unable to release data from which it would be possible to identify individuals.

Do existing policies address the needs of operational data?

Operational data were thought to be of increasing value to the scientific community in the future. At present, Argo float data are freely available and an IOC/IODE project is being set up to address ship-board measurements which would be made available in near real-time. It was considered important to address the needs of handling operational data in project data policies.

Policing

It was considered to be the responsibility of the project data management committees to monitor adherence to data policies and report their findings to the SSC. They should also promote the benefits of data sharing and endorsement by the project. Any failures to adhere to the policy should be reported to the SSC and will be dealt with on a case-by-case basis. It was generally considered unnecessary to involve the funding agencies at this level. Incentives should be developed to encourage scientists to comply with the data policy.

Conclusions

It was considered to be beneficial to construct a template data policy, and a small group should work together to distil common elements from existing data policies (see template in meeting report). The template policy should address issues such as the timely sharing of data and metadata, data formats and archiving of the data. The template policy could then be adapted by SSCs for each project.

As we are all ICSU bodies we should adopt the principle of free and open access to data.

Session on Technical Aspects of Data Management (Rapporteur Bernard Avril)

Interoperability for a distributed data management structure is required to sustain the possible growth of the structure. A set of possible remarks were made:

- Metadata are to be standardized.
- Tab-delimited, csv, netCDF data format should be promoted.
- One technical forum to set up the data and information management practicalities is required at the start of the project.
- Semantics and units should clearly be agreed upon, and data “markup” dictionaries (vocabulary) should be mapped and ultimately merged.
- possible entries by PIs of keywords for data search and retrieval should be made easier, in order to promote voluntary submission.
- There is a multiplicity of different and similar initiatives and some merging of the efforts should be recommended; several systems of distributed data sharing exist and some interoperability efforts should be recommended, through a dialogue within the data management community (if it can be defined, identified). Technical issues are currently discussed in several fora such as SeaSearch, EarthSystemPortal, Marine-XML, IOC-SGXML initiatives (especially for the dictionary mapping)
- In addition, funding should be carefully distributed between advanced research in data management and the basic work in project data management.
- In developing countries, there is a technology gap that might prevent them from benefiting from the new technological solutions. The need for a proper Internet access should be promoted and alternate, less-technologically advanced media support should also be made available.
- Even in developed countries, the simple security of data at the PI level is not always insured. An estimated 5% is attributed to DM in project. If the E2EDM structure (with adequate IT support) is setup adequately, it should be easy to attract and gather internationally the individual datasets when no other support is available.

New challenges

Basic structures for data and information management should be implemented and promoted, for example, through IOC, IGBP-START, SCOR, UNDP, and regional supporting agencies

If the technical advances are to be shared globally, the basic structure should also be promoted and implemented everywhere, and this would also help the gathering of data in the developing countries

The management of socio-economic (and other non-geospatial) data should also be addressed, especially through the participation of the appropriate representatives in those fields, within the DMTT.

Operational/real-time data acquisition is in marked development and the data management structure is not always adapted adequately. The Argo system is at the cutting-edge in this matter and should help to imagine the future structure for operational / real-time, marine data management. One aspect to be developed is the capture system and the access to those new data for various users (e.g., applied research, modelling), since the acquisition part is more advanced. The capture of the data flow and the incorporation within the existing DM structure is the weak point, because the data flow increases much faster than the capturing and handling capabilities.

Session on Resourcing Data Management (Rapporteur: Casey Ryan)

Key recommendations were:

If you start with 10% of a projects budget devoted to end-to-end project management, you can be sure that you will be well resourced and the probably feed back some money to science.

This is part of a bigger issue of funding project management. Operating the metadata catalogue should be considered a core activity of the IPO, and represents a min of 0.5 FTE.

The WOCE/DAC model struggled for funding, and the biogeochemistry community is still to show the benefit of making global data sets publicly available. Synthesis happens, but not the public sharing. pCO₂ and International Ocean Carbon Coordination Project may offer a way forward.

- What proportion of project budgets is assigned to data management?
 - 7% for Rapid (Doug says this is a bargain) max 15% in NERC. Overbudget and then hand back to science
 - infrastructure costs not included
- What proportion of project budgets should be assigned to data management?
 - 7-10% in UK + infrastructure costs
- How can resources be secured for development of international data management infrastructure?
 - WOCE did, JGOFS didn't
 - biogeochemistry hasn't shown the benefit of assembling the data, thus funding DACs is unlikely.
 - Synthesis happens, but making data available after does not. But maybe this is changing?
 - WOCE was global and still struggled for DAC funding, this is probably not a good model for SCOR/IGBP
 - WOCE SSC encouraged proposals for DAC funding, especially in U.S. JGOFS didn't
 - IOCCP and CARENA offer some examples.
 - IPO resources – data management may be the lowest priority, but should be included from a project's beginning.
 - Co-hosting IPOs with share data management infrastructure is an option.
 - There is a bigger issue of project management. This should be considered a core

activity of the IPO, and represents a min of 0.5 FTE for the metadata catalogue.

- How can we address the problem of managing data from minor 'voluntary' contributions to projects?
 - solution - end-to-end project data management with some additional infrastructure
- Are there specific resourcing problems for developing countries that need to be addressed?
 - no

**Session on Data Submission to World Data Centres
(Rapporteur: Liana McManus)**

Getting project data to the WDCs

Questions to address:

Should submission of data to WDCs be regarded as a publication process?

Should we make datasets citable? Definitely yes. How do we implement this? How do we do peer-reviewed datasets? We need an editorial board and a journal venue...AGU and EGS have had no feedbacks...getting a journal is a lot of work but getting online journals may be more feasible. This should be done through partnerships between data centers and a scientific society. Data centers can provide support for the review process but may not have the scientific expertise...we need an editorial board (AGU and EGS) and have links with data centers. The editorial board sets up the review process and chooses the reviewers.

A cultural change that has to take place, such as "brownie points" for data set publication. Data reports traditionally have not been peer reviewed. How does this discussion relate to journal requirements for authors to submit the data on which their papers are based? The problem with this mechanism is that the paper is reviewed, but not the data set.

We may recommend that datasets be deposited in some recognized data centres as requisite prior to the acceptance of papers for publication. ICSU has not had discussions on this. ICSU is doing a review of the data and information activities of its organizations (ICSU Priority Area Assessment for Data and Information) and has not come up with its recommendations –they are more concerned about data access.

IGBP, POGO SSCs can provide part of the reviewing process. The tenure process may not recognize dataset publication as part of those to be rewarded brownie points. We may start to change the culture about datasets.

Once data enters the WDCs, they are merged and will not be retrievable in their native formats.

Geophysical Research Letters (AGU) has very restrictive policies in terms of ref citations (only 15). Go to AGU directly about this.

IMAGES: data reports are cited. The interest of the scientist in delivering data is recognition.

How do you prevent authors from parcelling out data into several datasets to get multiple points?

Metadata recognition: attach names and roles.

Do the WDCs have a role in data quality control and assurance? If not, who does?

WDC for Oceanography, Silver Spring: synthesis and quality control. Objective analysis.
WDC-MARE philosophy. Data quality aimed toward publications.
CDIAC: more quality control with smaller datasets
WDC: more archival

IOC published standards of quality control. The issue is who does the function.

WDCs have a role; data originators, data users. WDCs have the function to archive the data quality control process. WDCs have different mode of operations. However there are variants in infrastructure that determine the extent to which WDCs can provide this function.

Should WDCs deal exclusively with national-level data management infrastructure?

WDCs can “back end” data managers.

Should WDCs get involved in the provision of “project data management” services?

YES.

**If so, at what stage (when in the project) and on what scale (to whom in the project)?
Do developing countries have particular problems submitting data to WDCs?**

Individual scientists can deal directly with WDCs. But this can be dangerous. Between the developing countries and the WDCs, there can be a communication gap. What do you do with countries that have no NODCs?

If a country wants an NODC, IODE can assist. One approach is for IODE to review WDCs, then for it to identify NODCs that can act as regional nodes.

Countries with marine projects should have a national data coordinator. The scientists should be made aware of available support for data management and there may be better ways to communicate the availability of these linkages.

IODE is realizing that it has to bring other expertise to help deal with the array of data that needs to be managed. In this case, IODE will have to consider distributed modes of data management.

Can the WDC network cope with all the types of data we anticipate collecting?

No. Most WDCs will try to evolve and take new kinds of data if we involve them in the dialogue. IGBP, SCOR, IOC may help coordinate the dialogue with WDCs, the IODE, etc.

Should WDCs actively track and pull in data sets or wait them to be “pushed” in their direction?

WDCs as such do this on a national basis, national relative to their host agency. They accept data from other countries, but do track it within their national locations.

Tracking and pulling data depends on partnerships between WDCs and data managers. And the partnership between WDCs and IODE, the latter may be the one doing the tracking and pulling, for example the GODAR workshops.