

Data management for IMBER

The IMBER Data Management Committee (DMC) was officially launched during the 3rd IMBER Scientific Steering Committee which took place in Brest (10-12th May 2006). The DMC will be chaired by Raymond Pollard (NOC, UK). The SSC also appointed the IMBER Deputy Executive Officer, Dr. Sophie Beauvais, as the IMBER Data Liaison Officer (DLO).

Raymond and Sophie met with Roy Lowry at the British Oceanographic Data Centre (BODC) in Liverpool, UK on 3-4th August 2006, in order to seek expert advice on some of the Data Management issues (see the notes from this meeting below).

The first task of the DMC will be to develop IMBER data management policies and metadata guidelines for IMBER endorsed projects. The overall objective is to ensure creation of full metadata and access to, sharing of, and long term preservation of IMBER metadata.

The initial task is to appoint members of the DMC, which will be composed of observationalists, modellers, data specialists and the DLO. It is hoped that the DMC will start its discussions, mostly by email, in November. In the longer term, the DMC and the DLO will together 1) ensure that the strategy is well implemented, 2) ensure that IMBER endorsed projects adhere to the policies developed by the DMC and appoint or identify a data specialist, and 3) keep track of IMBER metadata and make them available to the global community, via the IMBER website.

Notes made following a meeting between Raymond Pollard, Sophie Beauvais and Roy Lowry at BODC 3-4 August 2006 – (by Raymond Pollard).

Next steps

In random order:

1. Determine membership of DMC
2. Agree duties of DLO
3. Create guidelines for IMBER projects
4. Agree terms of reference of DMC

It would be good if membership of DMC could be sorted quickly, because then members can be involved in email discussion of 2 and 3. My suggestion for a timetable is:

Suggest names for DMC and approach people and seek email SSC approval between now and IMBER Exec. Approve DMC membership at IMBER Exec.

Agree 2 at IMBER Exec, but allow for changes as the role develops.

Agree 4 at IMBER Exec late Sept 2006.

Take to the IMBER and IMBER/GLOBEC Exec the ideas arising from the meeting with Roy, to have some internal discussion before starting the DMC discussion. I want to proceed cautiously on 3, as we may need to be diplomatic in selling some of our proposals to the wider community. Thus I seek to appoint to the DMC people who will bring a range of ideas, not necessarily in agreement at the start.

Major ideas arising from meeting at BODC

What metadata can realistically be achieved?

There is currently a gap between what IMBER scientists would like from data management and what Data Managers can realistically achieve. Our proposed baseline for IMBER metadata will be Cruise Summary Reports (CSRs) for every cruise and Directory Interchange Format records (DIF) for every non-cruise IMBER activity. This will allow us to track everything done under IMBER auspices and who is responsible.

It will NOT allow a modeller (say) to automatically pick up all data of particular types from various distributed databases for a particular area (say). Surely this has been done already, you ask? We believe not:

- WOCE set up international data acquisition centres (DACs) for particular types of data. Thus CTD data can be obtained from the CTD DAC (at Scripps), current meter data from the current meter DAC. But the data for a single data type are held at one place, not distributed.
- JGOFS data of many types can be obtained from the JGOFS data centre at WHOI. But this only applies to data that are held at WHOI (all US data maybe, but not all international data), and much effort was expended locally to achieve this, both in obtaining the data and in making it available for user specified areas, etc. So again the data are not distributed.

To allow access to distributed data (not metadata) requires a level of metadata documentation that is unrealistic at present. It requires, for example, that the metadata contain considerable detail about the data themselves, such as range of x, y, z, t coordinates; also that data access controls (for computers at many institutes and in many countries) are highly sophisticated. During the lifetime of IMBER there will be progression towards this, which IMBER may encourage, but it is not realistic for a baseline.

Even the suggestions below (CSRs for cruises and DIFs for everything else) will require some negotiation. Indeed, even this distinction is undesirable, and it would be good to have a DIF for everything. This will require software development to convert each IMBER CSR into a DIF, which can then be held in a DIF metadatabase. This too will need negotiation and has resource implications.

Cruise Summary Reports - CSRs (i.e. ROSCOP)

CSRs are a widely used standard (google ROSCOP or “cruise summary report” to find details). The main exception is USA. However, even there, if data are submitted to WDCA at Silver Springs, the data centre will generate a CSR from submitted material. Surely everybody has to create a cruise report, and the data centre will create a CSR from that cruise report if necessary. We recommend that IMBER should encourage

production of a CSR for every IMBER cruise. The DMC will need to discuss the pros and cons of this proposal.

There are printed forms, or Word templates to aid the producer. “Why not enter information directly into an excel spreadsheet?” Roy’s answer is revealing. Basically, a scientist almost always does not use the “controlled vocabulary” carefully enough. Does “temperature” mean sea temperature, sea surface temperature, air temperature, ... to give a simple example. Is “RRS Discovery” the same as “Discovery”? My impression is that it is felt to be less work, in the end, for data centre staff to transfer cruise report or CSR information in Word onto their data base than for the scientist to attempt it. As the data person cuts and pastes, he also converts to correct vocabulary usage, abbreviations, etc. Staffing implications are obvious. Most scientists simply do not give the attention to detail (nor have the data specialist knowledge) to create the CSR correctly. However, at some data centres the CSR can be entered interactively on-line, so, as research ships gain 24 hour links to land, this becomes a possibility.

There are two CSR repositories. An EU project SeaDataNet (google it for details) has as an objective to streamline CSR entry to their database at the German DOD (Deutsches Ozeanographisches Datenzentrum). This will become the major route for EU countries.

ICES is the other repository, which will accept global data. IMBER should explore with ICES their willingness to take in IMBER CSRs (as ASCII files). There are links between these two databases.

Given that cruises will surely be the major sources of IMBER data, it is worth ensuring that cruises are well documented, and the CSR is the standard to shoot for. However examination of the ICES and SeaDataNet inventories quickly revealed substantial gaps in cruises that I know about, so even creating a timely and complete inventory of IMBER cruises will be non-trivial. These cruises **are** in the BODC metadatabase however, so the problem appears to be lack of resource to populate the international databases.

Directory Interchange Format - DIF

For anything that is not a cruise, the CSR is useless, and the alternative standard is DIF (google it for details). There are on-going, well advanced negotiations to develop an international standard ISO19115 (google it for details) which will replace DIF, but conversion routines will also be available, so stick with DIF for now. DIF allows a summary metadata record to be created for anything, though it is usually done at the project level. Thus each IMBER project can be required to create a DIF as a summary. A major advantage of a DIF is that there is already a global metadata base of DIFs, the GCMD (Global Change Master Directory - google GCMD) which could be used as a portal to IMBER data. GLOBEC use it (from GLOBEC web site click on “data” then “metadata portal”). For another example, google JCADM (Antarctic Data Management project), then click on “data management” then “AMD”. We should approach GCMD to request this. GCMD provide tools to help with the creation of DIFs.

Converting CSRs to DIF

It is undesirable to have to search different metadatabases (GCMD, ICES, BODC) to find IMBER data. In principle, conversion of CSRs to DIF should be possible, allowing all IMBER metadata to be found at the GCMD. I think this is highly desirable. Can this be

fully automated? What resource will be needed? These are questions the DMC must address.

Guidance, encouraging and training

When I started trying to find out how to create metadata using web searches, I quickly became frustrated at the lack of recipe book “do it this way” information. There was plenty of top down “metadata are good” stuff, but only jargon when I tried bottom up. Talking to Roy has helped enormously, because he has quickly suggested recipes to follow. So even a willing scientist is frustrated when he tries to learn how to create metadata. I think the data centres are partly to blame, by not trying to educate scientists, but I detect that they are still trying to get their own standards in order, and lack resources even to keep up with metadata themselves.

IMBER can play a valuable role in several ways. It can push for some advances in marine metadata standards (say: universal CSRs, automatic creation of a DIF from every CSR). It can encourage development of tools to make the creation of some types of metadata (CSRs) easy for the scientist to do, with interactive error checking to ensure the metadata are not riddled with errors. It can educate marine scientists to the value of metadata and help the willing ones to create metadata by providing suitable recipes and tutorial material. For a start, the DMC needs to work with the DLO to create useful guidance on the IMBER website. This must be carefully done to encourage scientists, and show them that it is in their interests to play their part in creating accurate metadata.