Report on the symposium "Colour of Ocean Data" (*June 2003*)

Edward Vanden Berghe
Chair, Organising Committee COD
Manager, Flanders Marine Data and Information Centre
Flanders Marine Institute, B-8400 Ostend, Belgium
<wardvdb@vliz.be>

The 'Colour of Ocean Data' symposium was organised from 25 to 27 November 2002, in the Palais des Congrès in the centre of Brussels, by the Flanders Marine Institute, the Intergovernmental Oceanographic Commission of UNESCO, the Office of Scientific, Technical and Cultural Affairs of the Belgian Government, and the Census of Marine Life. Nearly 200 participants were registered; there were 44 oral presentations, 40 poster presentations and eight demonstrations.
The objective of the 'Colour of Ocean' symposium was to bring together different communities with an interest in marine sciences and information management. Along one divide, participants from the physical oceanographic data management were invited, as were the marine biology data managers. Along a second divide, marine/oceanographic data managers were confronted with the user communities, mainly scientists and policy makers.
In a series of five sessions, various aspects of data management were discussed. The main aim was to allow the different communities to learn about developments on related fields, and to learn from each others' experiences. For each session of oral presentations, there was a corresponding session with poster presentations and demos. The wide variety of topics that were discussed was indicative of the breadth of the field; the ensuing discussions clearly demonstrated the timeliness of the symposium.
The two last hours of the symposium were devoted to a short panel discussion. Two representatives each of international organisations, of the data management community and of the scientific community were given the opportunity to expand on their views on oceanographic data management, their views on the role of data centres, and expectations from user communities. The conclusions from this panel discussion will be included in the proceedings to the symposium.
One of the main conclusions from the panel discussion was that symposiums like this one were needed to strengthen the communication between different communities involved with marine/oceanographic data management. Some very practical suggestions were made. One recommendation was to make data sets citeable, so that scientists' motivation to submit data would be increased. A very strong recommendation was made to investigate how data management can be included in the curricula of universities. Last but not least, there is a need for data centres to be more service-orientated, taking an example from active libraries, rather than being grey and dusty archives, where data enters never to be seen again.

An extensive summary of the discussions during the panel session will be published in the proceedings of the symposium. A preliminary draft is attached to this document.

COD Panel discussion – Organisation and panel members

Objectives: Identify what data centres see as user needs and what users see as user needs.

In the context of ocean data management, scientists, data managers and decision-makers are all very much dependent on each other. Decision-makers will stimulate research topics with policy priority and hence guide researchers. Scientists need to provide data managers with reliable and first quality controlled data in such a way that the latter can translate and make them available for the decision-

makers. But do they speak the same 'language'? Are they happy with the access they have to the data? And if not, can they learn from each other's expectations and experience?

There were two panel members from each of the data management and the scientific communities, and from international organisations. The panel discussion was divided into two parts; the first part consisted of short opening statements by the panel members, based on the questions listed below. The second part was dedicated to open debate.

Questions:

Data Centre representatives:
(1) What do you see as the role for data centres in managing data from the global science programs?
(2) What are the challenges you see that data centres need to face individually and collectively?
(3) What added value comes from managing data in data centres, rather than in the originating institutions? What should a data centre have on offer to be more than just a convenient data archive?

Scientists:
(1) What are your expectations from the global network of data management systems?
(2) Can the global network meet your expectations now, with some changes or with radical changes?
(3) What governance structure would ensure effective and efficient management of global data, assuring and documenting data quality, securing data for future generations, and providing easy access to integrated multi-disciplinary data?

International organizations:
(1) What is the role of international organizations to address the data management requirements?
(2) What do you think are the major challenges that the international organizations face in global data management?
(3) What changes should be implemented at the international level to better deliver the global data management mandate?

All:
(1) What data management practices can be employed to reduce the impacts of technological differences between developing and developed countries?
(2) What do you see as the main differences in data management practices between biological and physical oceanographers? What can be done to bridge these differences?
(3) If you had three wishes to improve global data management, what will they be?

Panel Members

- Chair: Savi Narayanan, MEDS, Canada
- Representatives from data centres:
    - Lesley Rickards, BODC, UK
    - Catherine Maillard, IFREMER/SISMER, France
- Representatives from the science community:
    - Peter Herman, NIOO/CEME, The Netherlands
    - Neville Smith,
- Representatives from international organisations:
    - Alan Edwards, EU
    - Peter Pissierssens, IOC/IODE

COD Panel discussion - Main themes

Changing role of data centre

Changes in technology have been leading to changes in the role of data centres. There is a trend to move away from the traditional data centre, with its main task of archiving data sets, to become more service-orientated.

Data centres can look towards libraries for inspiration to redefine their role; libraries provide expertise and guidance in cataloguing. Archives are grey and dusty, libraries are active and open; data centres should strive to resemble the latter rather than the former. Data management needs an equivalent to the 'Web of Science': a mechanism to bring up a list of relevant, available, quality controlled, peer-reviewed data sets.

There is a need to create data and information products; not only towards other data managers and scientists, but also to the policy makers and society at large. These products will assist in increasing the visibility of the data centres, and so assisting in attracting both funding for further activities, and data submissions from scientists.

Some traditional roles of data centres remain important: long-term stewardship of data, integrating data sets, documenting and redistributing data sets, development of standards and standard operational procedures…

Bridging the gap between scientists and data managers

Both data centres, and data and information management procedures are very poorly known by marine scientists. In most university programmes, there is no training on data management, no information on data centres, data management procedures… Data management is perceived too much as an IT topic. There is a need to investigate how to put data and information management on the curriculum of academic institutions. This would result in a better knowledge of the data centres, and an increased quantity and quality of data submissions

Data managers should actively seek collaboration with scientists. If data managers have a background in science, it is possible to establish a relationship of trust with the scientists, a smoother collaboration, and a greater input of the data managers in the development of data collection. The involvement of the data managers in the planning of projects from a very early stage makes 'End to end data management' a reality.

EU has the mandate and the funds to support and improve training for scientists in data management, and could be playing a role in this.

Creating incentives for scientists to submit data to data centres

To a large extent, data centres are dependent on scientists to submit data. Especially in view of the extent to which scientists are not aware of the role or even the existence of data centres, this is a potential problem. Several actions can be taken in this respect.

- Creating awareness about importance of data management, by *e.g.* including data and information management in the curriculum of universities.
- Requirement for data management written into project condition for funding – is already the case for EU proposals, and happens for short-term data management.

- Developing peer review and quality control procedures, to assess usefulness of a dataset, and making a dataset citeable, so that a scientist's contribution of data to a data centre can be measured, and taken into account for career advancement.

Need for long-term activities

Data sets often result from projects, which usually have a limited time-span. Data management on short term, within the time span of the project, is usually no problem: scientists do need data management to produce the deliverables to the project; moreover, making provisions for data management is a prerequisite to have a proposal accepted in the first place. There is an obvious need for activities beyond the duration of the data-generating project, to assure continued availability of the data. This always has been, and probably should remain, one of the tasks of data centres.

Funds for long-term data management should not come from research budgets, but rather from operational networks or other mechanisms. Several initiatives of the EU are relevant in this respect. Within Framework 6, there is a possibility to fund the operations of large 'Networks of Excellence' that will operate on time spans much longer than a typical project. The Global Monitoring for Environment and Security (GMES) initiative is another potential mechanism.

Duplication of efforts

A certain degree of duplication is unavoidable, and is a fundamental aspect of the scientific process. There has to be room for experimentation, different attempts at solving the same problem. After some time, however, experimenting should stop and be replaced with one or a couple of strategies.

Undesirable duplication can partly be stopped during the project proposal review process. One of the objectives of the Networks of Excellence, as proposed by the EU, is to increase communication between partners of the network, raising awareness of each other's activities, and hence decrease the probability of duplication.

Need for peer review of data sets, and for standard practices

There has to be peer-review, as a way to measure and recognise progress, to recognise value and expertise, and as a foundation for standards and accepted procedures. Standards and audit procedures are needed to allow objective peer review. Developing these standards is a task for the data centres.

Peer review is a way to increase the compliance with standards. Countries, or even institutions or scientists, could be tempted to work along principles that were developed locally; obviously, these will fit local needs, and are usually much faster to develop. Doing so, however, can lead to fragmentation, and hamper data exchange.

Difference between biological and physical data management

The problems of biological and physical data management are different: physics data sets are often high volume and low complexity; biology data sets are low volume but high complexity. Taxonomy brings a 5^{th} dimension to ocean data management.

The lower level of standardisation in biology makes importance of proper documentation with the data sets even greater.

Commonalities are more important than differences: both biological and physical data management need for long-term activities; need for quality control and peer review; need to create data products

Involving the developing countries

Participation of developing countries in global programmes is the best way to transfer expertise. Global programmes can operate at several levels, so that they can serve both global and local needs.

Internet access is a problem in many third-world countries, and assisting with connectivity and basic telecommunications should be made a priority in any capacity-building programme. Where internet is available, the bandwidth is often very limited, making it virtually impossible to download large volumes of data. As long as this problem still remains, data should also be distributed on alternative carriers such as CD ROM or DVD. Data warehousing and brokering can assist in locating and selecting relevant data sets, and thus limiting the volumes of data to be downloaded.

Also funds to purchase hard- and software, and expertise to maintain the systems, is a factor that is more limiting in developing countries. The data management community should provide platform-independent software that is open source and runs on hardware that is compatible with technological expertise available. Reliable and stable standards should ensure that data are available in a form that can be handled by these tools. Capacity building programmes should be organised making use of these tools and standards.

Actions

- Investigate how data and information management can be made part of the curriculum of marine sciences in academic institutions
- Develop standard operational procedures and a peer review process to allow an objective assessment the quality of data sets
- Guide the user community directly to relevant, and quality controlled data sets, by setting up portal sites
- Create integrated data products, to increase the visibility of the data centres
- Distribute data not only through the internet, but also on CD or DVD
- Assist third-world countries with basic telecommunications, internet access, and data warehousing
- Create a collection of open source, platform-independent software for the benefit of third-world countries, and organise capacity building programmes around these