

Appendix II Presentation Summaries

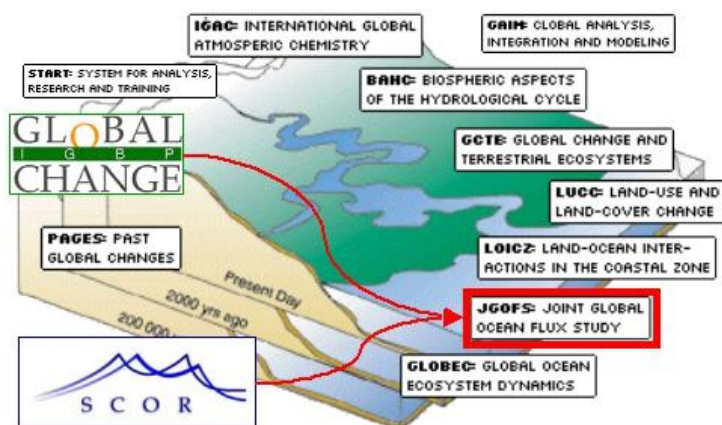
JGOFS Data Management -- Bernard Avril

Marine Data & Information Management (D&IM) – Lessons learnt from JGOFS Bernard Avril (JGOFS IPO)

This presentation is written with the overall idea that the **highest quality in D&IM** should help and support the PIs, project sponsors, funding agencies and end-users, to achieve and/or benefit from the **best science**, today and tomorrow. *N.B.* This presentation is not always (or ever) the “official” viewpoint of the JGOFS community, if such thing can be defined... Yet it is a viewpoint, and it could be sometimes useful, and discussed...

Science & Data from *Joint Global Ocean Flux Study: What is JGOFS?*

The Joint Global Ocean Flux Study (JGOFS) is an international and multi-disciplinary programme with participants from more than 20 nations. JGOFS was launched in 1987 at a planning meeting in Paris under the auspices of the Scientific Committee of Oceanic Research (SCOR). Two years later, JGOFS became one of the first core projects of the International Geosphere-Biosphere Programme (IGBP).



The JGOFS Scientific Steering Committee (SSC) created a number of Planning Groups and Task Teams to consider scientific and logistic questions and make recommendations. These groups helped to identify and plan the most important processes and variables to study, the ocean regions that such studies should provide the greatest insight and most useful data, and the best experimental design for the studies. The SSC set up the sequence of events necessary to complete specific tasks, the resources and level of international coordination

required for the tasks. The following scientific goals of JGOFS were published in its Science Plan to determine and understand on a global scale the processes controlling the time-varying fluxes of carbon and associated biogenic elements in the ocean, and to evaluate the related exchanges with the atmosphere, sea floor and continental boundaries, to develop a capacity to predict on a global scale the response to anthropogenic perturbations, in particular those related to climate change.

The strategy for addressing these goals has included a series of process studies in regions of the ocean that are thought to contribute the most to the flux of carbon between the ocean and the atmosphere, a global survey of dissolved inorganic carbon parameters in ocean waters, and several long-term measurement programs at sites in key ocean basins. JGOFS developed a plan for synthesizing the observations into a global picture of large-scale fluxes with the help of several modelling techniques. JGOFS was indeed committed to the development of models that can assimilate results from field studies, produce accurate large-scale descriptions of ocean biogeochemical phenomena and predict oceanic responses to environmental changes. The final component of the JGOFS strategy was a comprehensive and accessible database of results.

JGOFS has completed more than a decade of field studies in key regions of the global ocean. These studies have brought together data on chemical fluxes, biological processes and the physical forces that constrain them, in order to increase our understanding of the pathways by which carbon moves through the ocean in various forms, organic or inorganic, in particles or dissolved in the water, and of the ways in which biogeochemical systems vary over time and regionally. Finally, the availability of

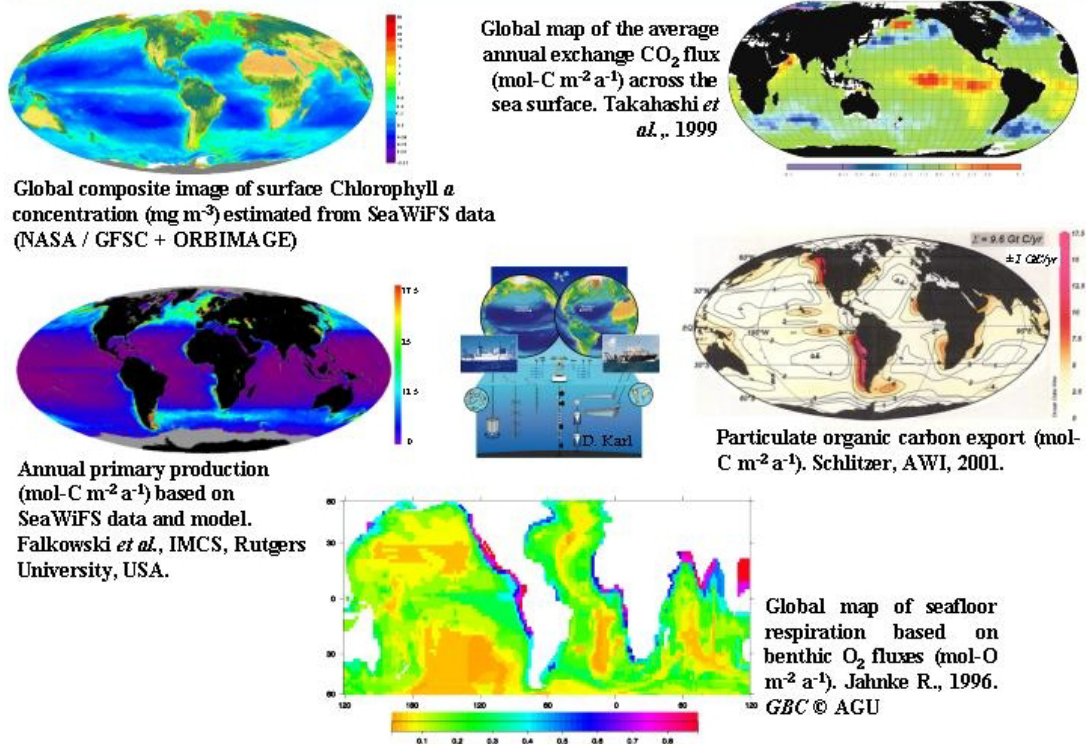
remote-sensing data from instruments on satellites is making it possible to extend the inferences made from the JGOFS field studies to regional and global scales.

The JGOFS science produced an **unprecedented, high spatial & temporal resolutions of complex, multidisciplinary ocean data** (with more than physical, chemical, biological and sedimentological core parameters), acquired through a set of (evolving) extensive & intensive, marine biogeochemistry studies, helped the design of field projects, hypothesis testing and diagnostic & prognostic models, and required an adequate data management (& information) system

According to the JGOFS Science Plan (1990) and JGOFS Implementation Plan (1992), **the initial objective and proposed set-up of the JGOFS Data Management System** were as follows:

The goal of JGOFS Data Management System is “to provide all interested scientists with *complete and convenient access* to the international JGOFS dataset.”

The **JGOFS policy** regarding data submission, access and exchange, encourage free & open communication of findings, including raw data, recommended access without restriction for any



interested user, provided that data originator is contacted for permission of further data use, encourage national (*mandatory*) submission by project investigators of their data to the project data system in a timely manner (cf. core parameter protocols), and invited each national project committee to endorse this policy.

The National Project Data Centres (with a data coordinator) were expected to provide flexible, online procedures of submission, exchange, retrieval & work, to establish an online national project database, including data inventory & metadata, to use (extra-)national exchange modes and formats for data project centres, to arrange training in using the procedures, and to interact with the project DMTT as needed.

Data Management Group (later *DMTT*) was expected to continuously co-ordinate the national activities, as needed, to monitor data management compliance, and to reassess periodically needs & performance.

The **Core Project Office** (later *IPO*) also involved in the project data management, with the goal to help the investigators to locate and access project data, through the support of project planning & execution, data management, synthesis & modelling, and as a project resource centre on national and regional activities.

The **Terms of References** (according to their last revision in 1997) of the **JGOFS Data Management Task Team** were to initiate, encourage and develop national data management, to develop, implement timely, national data exchange mechanisms, and international data integration, to compile & disseminate an integrated, international data and cruise inventory, to monitor international acceptance of, compliance with, policy (& adapt it), to liaise with other (inter-)national initiatives & data centres, to provide IPO with advice regarding data management, and to report to SSC and implement its recommendations.

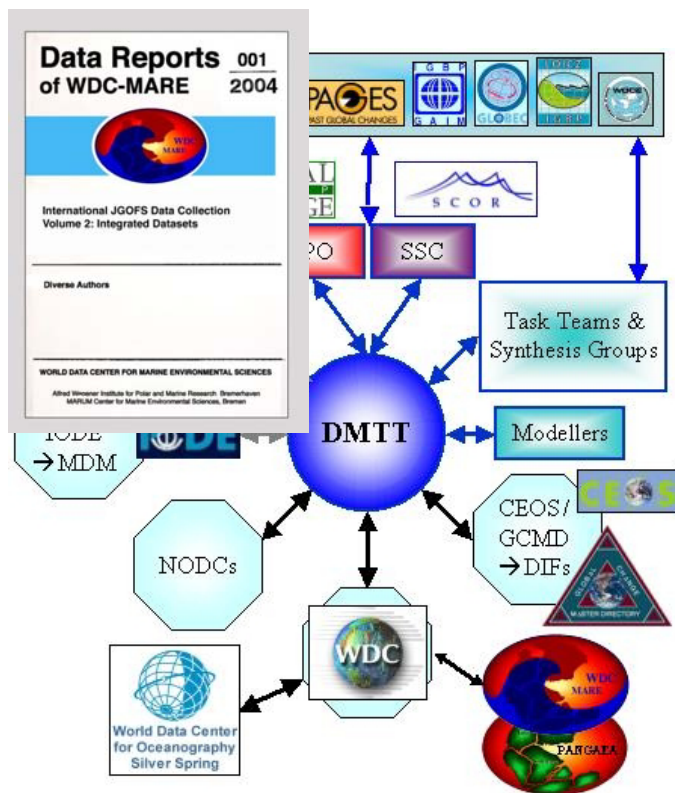
The DMTT was composed to several national representatives, located within their NODC (CA, IN, JP, UK), or in an oceanographic institution (AU, FR, GE, US). In other participating countries, data mostly with individual PIs (e.g., IT, PK, SP) and rarely managed nationally (NL, NO).



According to the Final Report (2003) of the JGOFS Data Management Task Team, the main achievements were to provide a high profile for biogeochemical data management within community and outside, to compile and publish national cruise and data inventories, with added value, to produce a unique, multinational, biogeochemical data product (DVD vol. 1), to contribute to WDC-A, NASA's GCMD, WDC-MARE, and other initiatives. The DMTT also contributed through the publication of multiple CDs, the management of online data repository, and with the help of the IPO also contributed to publications, special issues, reports, meetings, books

The **JGOFS International Data Collection** is gathered in a

“Volume 1: Discrete Datasets” DVD which is the first large scale, international, discrete data collection in global oceanic biogeochemistry, from about 1000 cruises, and is a fair representation of the national JGOFS data managers' efforts. It represents from 80 to 96% of the JGOFS-funded science data available, from the DMTT-represented countries. A rough calculation indicates that about 85% of datasets or cruises inventoried in the JGOFS DVD vol. 1 are originating from DMTT-represented countries. According to the last updated JGOFS cruise inventories, about 80% of all JGOFS data are on DVD vol. 1. In addition, a rough estimate of the funding dedicated to data management is about 5% [2-8%], excluding the approximate 5-yr half-time IPO support

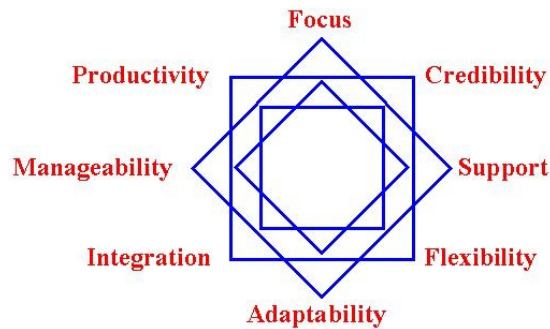


It is anticipated that a JGOFS International Data Collection. “Vol. 2: Integrated Datasets” DVD will be prepared with the help of the WDC-MARE, funded by AWI and BMBF (Germany). So far, about 40'000

JGOFS data entries are already available in the PANGAEA system (www.pangaea.de/)

According to the Final Report (2003) of the JGOFS Data Management Task Team, the “**Lessons Learnt**” from the experience gained in the JGOFS Data Management, and which could serve as

recommendations of improvements for the benefits of all parties for future Data and Information Management (mostly, to ensure the rapid dissemination of data and its long-term preservation and accessibility) could be summarised as follows: Establish a coherent, credible, semi-distributed and scalable, end-to-end D&IM plan with clear D&IM objectives, activities and timeline; a policy (e.g.,

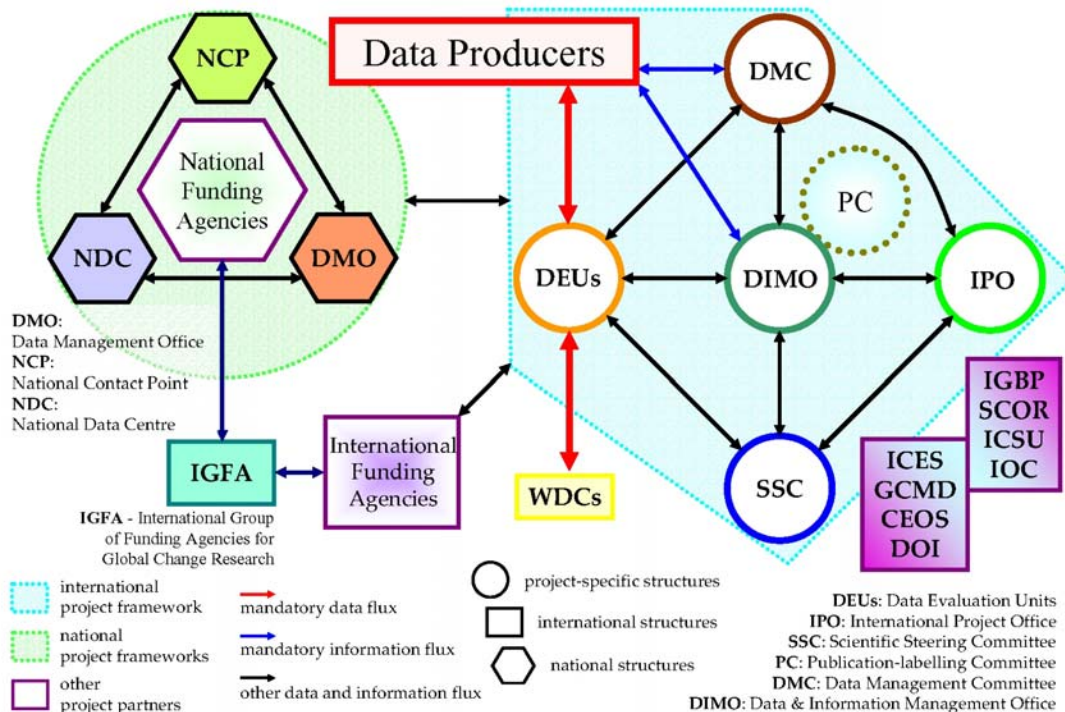


delivery and exchange standards), widely agreed by the scientists and sponsors, and supported by funding agencies to insure compliance; a clearly identified core project science and associated core parameters, to be revised regularly; an International Project Data Centre designed to establish guidelines, provide advice, and facilitate exchange of knowledge and expertise; several experienced, full-time national data managers / coordinators; an increase of the overall value of scientific research and derived outputs...

In addition, it is proposed some additional ideas for a better Data & Information Management, through a set of issues and questions to be addressed through few keywords describing the overall D&IM system:

- **Focus:** definition of the project scope (*business plan?*): e.g., identified project (inter)-national activities, core parameters, expected data volume and data management & product requirements
- **Credibility:** Coherent & credible data management framework: "An attractive carrot is worth ten big sticks"
- **Support:** full and adequate support of the PIs, project managers, project sponsors and other end-users..., and reciprocally!
- **Flexibility:** levels of granularity, of stability and of efficiency: e.g., DM organised by PIs, projects, nations, parameters, or scientific communities?
- **Adaptability:** roles of already existing N(O)DCs? e.g., long-term archival / on-demand, multi-disciplinary, multi-strategic project support
- **Integration:** added-value for all integrated activities and data product, for all parties: e.g., no duplication of works, standards, easier synthesis and modelling
- **Manageability:** Objective tools to review, evaluate, monitor the D&IM efficiency and achievements: e.g., data delivery, data usability, capacity building.
- **Productivity:** create highly visible, international deliverables to maintain the D&IM dynamics

An **idealised D&IM system** is also proposed as the support for a possible vision or a feasible set-up



for the future D&IM for marine projects.

It includes a range of services and products, taking full advantage of best practices, standards or innovative approaches; several pro-active, "bottom-up" strategies addressing all needs and requirements; the establishment & support of experienced, full-time national or regional data managers, through existing infrastructures, capacity building, new international synergies; a fair "top-down" strategies to insure compliance, better integration; a policy fully agreed upon and widely disseminated at the (inter)-national levels, for delivery, quality-control, referencing, exchange, integration, dissemination, preservation..., to facilitate interoperability; project knowledge resource centre (DIMO), a Data Management Committee (DMC, with data managers, observationalists and modellers) and a set of Data Evaluation Units (DEUs), where needed.

Within the schematic diagram representing the interactions between the project partners and collaborative bodies, and the main fluxes of data & information, each element represents a function rather than an office.

In addition it is reminded that several useful marine data policies are already available (especially online). A rough compilation of unofficial comments from data managers is also available.

As indicated by H. Ducklow (JGOFS SSC Chair) during the Final JGOFS OSC (May 2003), *JGOFS* was created as a visionary programme, especially through its union of biology, chemistry and physics, (*and sedimentology*) and its collaborative science works including more than 20 nations, in a global, marine *biogeochemistry* (V. Vernadsky, "*The Biosphere*", 1926) project. This is a major achievement of *JGOFS*.

And behind all those science works and approaches, there is also a revolution in the way data and information have been managed! ... *and it is not finished!*

WOCE Data Management -- Jim Swift

WOCE Data Management

James H. Swift
UCSD Scripps Institution of Oceanography
08 December 2003

This document addresses "WOCE data management" as opposed to "WOCE Hydrographic Program data management." This is a subject of truly enormous scope, considering that WOCE addressed a huge range of measurement types and was intensely multi-national. Apologies are made for any misconceptions that are inadvertently perpetuated due to the narrow focus of the author's prior work in this area.

This document begins the author's replies to a list of questions provided. Following these replies is a discussion of additional issues relevant to the theme of WOCE data management.

A separate document attends to the subject of WOCE Hydrographic Program data management.

What were the basic features of WOCE data management?

The World Ocean Circulation Experiment (WOCE) during the 1990s included several large measurement efforts of global scale. In the WOCE scheme of data management, all data were handled by data offices, usually what were termed "data assembly centers" (or DACs).

The basic features of WOCE data management included utilization of data assembly centers as intermediaries between measurement groups and the principal archive centers, scientific oversight of the data assembly centers and their activities, coordination between data assembly centers, and creation and utilization of a Data Information Unit which was a cross-cutting highly activist information center providing to and receiving from investigators and data centers a wide range of information about WOCE and WOCE data.

The official rationale for the establishment of the DACs is unknown to the author, but in a practical sense in order to permit analyses of large-scale and/or which utilized many different data there were needs to (1) move data as efficiently as reasonably possible from the data originators to major data centers, and (2) to make it possible to use disparate data together. Specialist data centers with direct involvement by practicing scientists, with scientific oversight bodies as needed, was seen as the route to follow. There were many good intentions and several attractive concepts. All in all, the system worked, though in hindsight the entire DAC process was significantly influenced by human nature.

WOCE for the most part did not create new measurement types. Therefore to some degree there existed an infrastructure for handling the data, i.e. many of the DACs already existed in one form or another. But even at the first stages of WOCE planning it was realized that something more would be needed in nearly every case, and that a few new pieces of data infrastructure were required, and most existing DACs needed augmentation of original activities and funding.

This is a nearly complete list of WOCE Data Assembly Centers:

- satellite (mostly SST & SSH; both at the JPL PODAAC)
- upper ocean thermal (XBT) - three separate DACs
- CTD/hydrographic/tracer
- current meter
- floats (subsurface)
- drifters (surface) - two DACs
- surface meteorology
- ADCP
- sea level (tide gauges) - both "fast delivery" & "delayed mode" DACs
- surface salinity

- Data Information Unit - master information center

The Science Steering Group determined which measurements were 'WOCE measurement's. Hence, for example, although there were usually CO₂ measurements on WOCE Hydrographic Program One-Time Survey cruises, CO₂ parameters were officially JGOFs measurements. The WOCE DACs which handled the companion data from these cruises - the WHP Office (WHPO) - merged the carbon data with the other parameters and worked with the US CDIAC to see that the WOCE data files contained the appropriate version of the carbon data.

How were data managed and quality controlled within WOCE?

WOCE data were managed by Data Assembly Centers each restricted to a specific type of data. The quality control process differed somewhat depending on data type. The author does not know the procedures used by DACs other than the WHPO.

The Hydrographic Program may have been unique in its QC structure: The procedure laid out pre-WOCE by WOCE planners was that data originators would assign quality codes (as per a WOCE-sanctioned scheme) for each parameter. They would submit the data and documentation in final form to the WHPO according to a [very optimistic] time scale. The data and documentation then were to go to external Data Quality Experts (DQEs) to have a second quality code assigned and make other recommendations about the data. In the original scheme the data then were to go back to the Chief Scientist, who would reconcile the differences and respond to the recommendations. This rationale for the extra, external QC procedure was likely drawn from the experience of the few most experienced compilers of pre-WOCE global data sets, who reported that many hydrographic data required adjustments (i.e. corrections for calibrations) and excise of bad values.

For the WHPO, the actual process was flawed on several levels:

Reliance on the Chief Scientist: most WHP Chief Scientists were tied to a specific measurement program and had no time or funds, and little inclination, to manage the wide range of other data from their WOCE cruise, taken often as not by groups the Chief Scientist was happy to have along at sea but with whom there were few interactions. WOCE needs were often stated as "requirements" and these did not always fit well with the needs of the seagoing science team, which had more immediate goals. Seagoing science teams rarely stayed in close contact with the Chief Scientist after a cruise.

Time required to provide final data: There was a huge range in the time taken by seagoing groups to provide final data, in almost every case far longer than the WOCE timeline. This range covered the groups at sea on one cruise as well as cruise-to-cruise differences. The gap lay partly in the difference between usable data and final data, the time required to prepare documentation, and the effort to assign WOCE quality codes. The delays were compounded by most scientists' needs to move on to other projects. And many scientists preferred to retain data privately until they and/or their students had published work on the data.

Shortage of Data Quality Experts: Most of the known expert persons thought ideal for examining WHP data were continually 'interrupted' in their DQE chores by their principal employment. Although not the most serious issue, this did limit the final progress on the WHP DQE tasks.

It was realized, however, that many scientists and data groups did a credible job of internal data quality control, spurred on, perhaps, equally by WOCE data quality needs and peer pressure. The WHPO began to work more closely with the individual data providers (e.g., the CFC group at sea). The external DQE program was assessed and prioritized. And community groups took on a few of the parameters (e.g., CFCs). This last step was a major improvement over the original concept. It was also realized that some of the WOCE Atlas groups were in fact carrying out a superb point-by-point data examination and were communicating their findings of questionable data to the WHPO. This was an outstanding substitute for external DQE for the ocean regions covered by those groups. [If public acknowledgment is appropriate in this document then the praise should go to Lynne Talley (UCSD/SIO) for her extraordinarily thorough discovery and reporting of WHP data questions.]

How did WOCE data enter NODCs and WDCs?

One of the functions of each WOCE DAC was to see that their data were transferred to the principal data centers. Some WOCE DACs - the ones which existed prior to WOCE - had long-standing arrangements for archive of their data. But during WOCE the three WOCE-wide data compilations also served as a primary means to see that WOCE data entered the archives.

NODC's strategy for archiving WOCE data was relevant to supplying data to users and insuring against loss of information into the future, rather than, for example, integration or search activities. The onus was on the DACs to prepare data, and the WOCE Archive simply preserved what the DAC submitted. NODC, in its archive of any submitted data, assigns the submission to an "Accession", defined as a retrievable collection of files associated with searchable metadata, for which NODC will monitor its integrity and perform future media migration. In the case of WOCE, each DAC's submission was to be given a separate Accession identifier, one clearly recognizable as of WOCE origin.

Some WOCE data were to be ingested into the NODC databases as well as being provided whole via the accessions. These included Upper Ocean Thermal data which go into the GTSP database, and the CTD/hydrographic data which go into the Profile Database. Surface drifter data are archived at the RNODC at MEDS. Other data did not fit into the standard NODC databases (time series, current meter data, etc.) and so these data will remain only in the WOCE Accessions.

The author notes that the WHPO and NODC were successful in mapping the WHP QC flags to the standard IGOSS flags used in the NODC Profile Database. This ensured that the most essential WOCE QC information will remain with the data when they are ingested into the NODC database.

What percentage of the WOCE total budget was devoted to data management in the most recent year?

The author apologizes that the scope of this question is entirely outside his experience. The following first-order estimate is for the WOCE Hydrographic Program only, and even this is more or less an informed guess.

\$ 30,000 daily science team cost for a One-Time Survey cruise

\$ 20,000 daily ship & resident technician cost for OTS cruise

\$175,000,000 100 One-Time Survey cruises at 35 days each

\$ 17,000 daily science team cost for a Repeat Hydrography cruise

\$ 15,000 daily ship & resident technician cost for Repeat cruise

\$307,200,000 400 Repeat Hydrography cruises at 24 days each

\$482,200,000 Total Cost of WOCE Hydrographic Program sea program

If the total NSF expenditure on the WHPO over 12 years was about \$5,400,000 (in the same type of dollars) with another \$3,000,000 (???) for the WHP-related duties of the WOCE DIU (and the WOCE Special Analysis Center [otherwise not covered in this report]), then $\$8,400,000 / \$482,200,000 = 0.017$ or 1.7% of WHP money was directly spent on data management.

How were the WOCE SSG and IPO involved in data management?

WOCE was a large enough enterprise that the SSG played only an indirect role in data management, basically setting priorities and policies into play, to be enacted and realized by various standing committees. The WHP, which was the largest program overall, had its own scientific oversight committee, the WHP Planning Committee, during the pre-WOCE years and up through about half of the WHP. The Hydrographic Program was alone in WOCE in the earlier years with its own scientific

oversight committee. This was necessary due to the scope and complexity of the program. The other WOCE DACs were overseen by the WOCE Data Management Committee. As WOCE matured, the need for WOCE-wide integration was realized, and the DMC evolved into the WOCE Data Products Committee. For a brief period the DPC and WHP-PC co-existed, but the WHP-PC eventually dissolved.

The WHP-PC was very highly involved in WHP data management, and the WOCE DMC/DPC was very highly involved in WOCE data management. These groups met as often as needed - at least once per year but more often in the earlier going - providing advice and encouragement to the DACs and carrying special concerns to the WOCE SSG. Official sanction was essential in obtaining fiscal and personal support for WHP activities, including data management activities. The close relationship of official planning bodies to the DACs was thus mutually beneficial.

What lessons has your project learned about data management?

The most important documentation ("metadata") should be permanently attached to the data. Too often data become separated from key information. Every data file should ideally contain sufficient information, whether by direct content or by obtainable reference, to understand and use the data. This concept looks great on paper but is quite difficult to implement.

Documentation is the key to long-term service life of data. There are, in one way of thinking, three key elements to reference-quality data: documentation, documentation, and documentation. This is well known, of course, but bears emphasizing.

Data management is too important to be left to data managers. If the end goals of data management are to support both data originators and data users, then active scientists and technical teams from both must be involved in data management and the oversight of data management.

Be careful what one asks of data providers. There are aspects of data management facilitated by imposition of requirements upon those providing data to the data centers. If data originators adhere closely to these requirements, the tasks of the data center are understandably simplified plus the data are more likely to fit into archives and in other ways have a long service life. That said, it is clear to the author that such requirements can at times be unintentionally overbearing, and for example can represent a harsh burden to developing technical teams and/or to teams for whom the language of the requirements is not their first language. Yes, some data centers may spend 90% of their effort on 10% of the data, but that is part of the point of having the data center, especially if otherwise valuable data would go unreported. The approach of working more closely with data originators who are having difficulties, keeping in mind their cultural and technical origins, seems more wise than raw listing of pages of requirements for data reporting.

Human nature is not changed by issuance of guidelines and requirements. The American Geophysical Union has very strict guidelines for data availability for research articles in its journals (available on the AGU web site). Yet it appears to the author that these are routinely ignored by authors, reviewers, and AGU editors. Even the U.S. NSF rule that NSF-funded investigators must make their data public within two years of measurement has proved difficult to enforce. Therefore, oversight bodies and data groups may publish all the data sharing, submission, content, format, and documentation guidelines they wish, but these should account for what is known about human nature rather than attempt to thwart it. What can a data center offer a data originator in return for hard-won data?

Maintaining data pedigree is crucial. Keeping accurate, easy to read records of each transaction involving a particular data set is crucial. Metadata should be inseparable from the data themselves whenever possible in the form of succinct data histories and version control numbers embedded in the data files. An EOS article from a few years back covered the unfortunate circumstance of a scientist who wrote a research paper based on dated (and incorrect) results. He needed the data that had been subsequently calibrated instead.

The WOCE Data Committees and the DAC System were strong assets to WOCE. The WOCE DACs benefited from the major and minor course corrections provided by the other DACs and the oversight

bodies. The oversight kept both the WHPO and the whole WOCE data system accountable in an objective way. The DAC system promoted the development of a common language and the opportunity to come to agreements on operational standards that each data center should use.

Consider what project-wide format specifications will be useful. The WHPO staff observe that JGOFS adopted fairly loose policy on data submission formats, but they absolutely required and enforced having certain elements in each submitted data file. Furthermore, and this is a major point, there certainly were no program-wide format requirements (or even suggestions) for WOCE between data areas. Of course, one cannot expect that the Upper Ocean Thermal DAC would include EXPOcodes in their GTSPF formatted files, but the fact that format compatibility was not discussed much (if at all) between DACs until very late in WOCE came back to haunt the WOCE data system during the technical data integration meetings. In other words, WOCE, as in the overall program, did not have any specific formats. Instead there was at least one for each data type or DAC with no opportunity to coordinate the contents of each disparate format.

Use simple off-the-shelf technology: Each program, especially data programs, should look at the day of their demise as a first-order exercise. No data group or center lasts forever, therefore the group should implement, whenever possible, use of standard operating systems, software and data format conventions. All too often, a group will use a system that is convenient for them, which many times will lead to confusion for those who follow them. Technological requirements should be continually examined to avoid both too-heavy immediate requirements or longer term potential dead-ends which may leave present or future data users without access to data.

A superb data manager: Canada DFO / Bob Keeley. The WHPO Data Manager finds Bob Keeley to be one of the best data managers: "He is well versed on related technical issues, but has no problem admitting that his vast experience is somewhat dated. He frequently makes his views heard, without offending other committee members. His views and opinions are managerially and technically sound (and feasible to implement). I recommend that this group solicit his views before making any final decisions."

What limitations would be faced in sharing WOCE data with other projects and what steps could be taken to overcome such limitations?

The author is not certain how to approach this question. WOCE data were managed for the long term as much for WOCE itself. Major emphasis was thus placed upon providing appropriate documentation and other data information, in improving overall data readability, in coordinating terminology between data assembly centers, and in ensuring data availability. In the short term focus was upon the measurement teams and immediate data users but the program never lost sight of the long-term users of WOCE data.

Discussion of Additional Issues Relevant to WOCE Data Management

WOCE Data Integration

An attempt was made to provide an integrated WOCE dataset by the time of the WOCE Final Conference. (The "Version 3" data set was due at this time.) This effort was not initiated by the DACs but by the scientific oversight bodies. The concept of integration had to do with joint use of disparate data types, for example locating and using all WOCE data from a certain region and time. WOCE Data Products Committee initial objectives for WOCE data integration were:

- 1) Define the scientific requirements for integration for V3 of the WOCE Data Resource;
- 2) Identify the level(s) of integration necessary to meet this requirement (e.g. is searching for data by expocode, time, and location sufficient?);
- 3) Evaluate relevant software/hardware technologies that consider a single and/or distributed WOCE data source(s) that satisfy (1) and (2);

4) Recommend a strategic plan for developing the integrated WOCE Data Resource.

A user survey showed that WOCE scientists wished to search all WOCE parameters in any number of combinations, including with QC information and metadata with them. The most requested linkages and selection criteria were position, time, availability of variables, QC flags, cruise labels and expocodes and data types. Most users were happy with netCDF for the data; however, a significant number said they preferred ASCII for small datasets and the long-term archive. Users were asked to provide their thoughts on static media (CDs, DVDs) versus online servers. There are pros and cons to both, and the outcome was that both are required for WOCE.

In the end the parameters the WOCE DPC agreed that should be in an inventory for data searching were Latitude range, Longitude range, Time range, Depth range, Variables (this category was hugely shortened in the end), Experiment Identifier (expocode/experiment name/station number), and Platform type (e.g. CTD, XBT, SL station etc).

The V3 WG decided that for practical purposes the static media integration would be at a lower level than any online integration, simply on the grounds that existing technology was driven by online applications. But it was also realized that there was no real need for the WOCE data to be on distributed servers once the V3 ("final") version was compiled. However it was recognized that some components of the WOCE data resource were part of ongoing projects and hence would continue to grow. In other words, if there were an online version of V3 it must evolve with other projects. Essential elements therefore were a continually updated inventory and commitments for support of ongoing on-line access for serving V3. Options for data delivery ('DODS', and 'CORBA'), GUI front end ("Live Access Server", "Climate Data Portal", and Matlab in DODS), data inventory and database ("LAS" or list-oriented files) were discussed by the WOCE DPC. Although there were several groups developing GUIs for search and retrieval of data online, none satisfied the WOCE requirements. (In the end, this author believes that a permanent solution for on-line servers was not reached.)

The static media options were much more limited. For example, the DODS and LAS server options would require that the user install software on their local machine. This was not considered practical because the software is not cross-platform compatible, installation is non-trivial, and search and retrieval procedures are impractical on "low-end" machines. The static media would, however, contain a searchable data file inventory, and the data themselves be in a consistent, self-describing form of netCDF. The DPC agreed that the most important next step was the development of the data file inventory tables to facilitate searching of the data DVD contents. (The author believes that this was the level of solution reached by WOCE.)

The requirement for the DACs was to generate netCDF files and data inventory tables, meanwhile maintaining the consistency of netCDF files between DACs, on top on their ongoing efforts to assemble and document the data for the WOCE Version 3 data set (and for the Archive). Not all DACs had resources sufficient to consider even these basic data integration issues. In the end it was decided that the design of the contents of the data file inventories should be carried out by the V3 working group, the creation of the tables by the DACs, and that the WOCE DIU would assemble the tables.

Additional comments: Essentially the data file inventory tables provide the core information that is needed for everything else; without them, nothing else would work, and with them, WOCE could aim for a range of integration levels from basic to complete. It was more practical to construct a separate table for each data stream. The tables would be flat ASCII, comma separated, with one record per "data unit" (e.g., file). There were to be a set of columns common to all DACs (generally expected to contain variables users will want to search on) and then a series of columns that would vary from DAC to DAC. The first row of the table may contain a description of the content so a search engine can immediately see whether the data files contain a particular variable, for example, temperature. It was necessary to devise tools or scripts for automatic generation of the tables from the netCDF files because the inventories will need to be updated as datasets develop. The tables in this form could be used for a range of activities from manual searches in a text editor, through basic search facility for the V3 CDs/DVDs, to complex search and retrieval online. The toolkit to create the inventory table was adapted from tools that existed to interrogate netCDF files.

Two principal components came together for the WOCE integrated data set: specifications for WOCE V3 netCDF data files and specifications for WOCE V3 inventory files.

A revised set of conventions for WOCE V3 netCDF files was developed, corresponding to the "COARDS" standard used in the model/forecasting community. Challenges for the WOCE DACs included: adding a new variable "time" with specified format, requiring certain variables and relevant attribute, and providing more explicit specifications for structure, variable names, and a range of attributes. The new time variable was the most difficult, yet the most critical to add. The DIU helped considerably to work on the problem of translating between WOCE date/time and the new time variable.

A critical component of the search/integration tools was the inventory tables that provide the critical parameter space for each WOCE netCDF file. Each WOCE netCDF file required an entry in those tables, which serve as the foundation for searching the WOCE V3 netCDF files.

The DIU developed an off-line search tool with a nice graphical user interface. This tool along with the inventory tables would also provide the fundamental underpinnings to query an on-line server, thus it may be possible for users to search WOCE V3 data on-line.

WOCE Data Orphans

There were several "data orphans" identified during WOCE, that is, data either unforeseen due to development of new technologies (collecting and processing lowered ADCP data and obtaining profile data from PALACE floats), or lack of mandate, awareness, and/or capability at a DAC to deal with a data type nominally in its purview (to some extent sea surface salinity and hydrographic data from mooring cruises). One of the roles of the DPC was to see that these issues were resolved. This was intertwined with the issue of uneven availability and support for the various DACs. In the end there was only so much cajoling that could be done, but little of direct WOCE interest failed to make it into the DAC structure.

WOCE "Reconciliation"

One of the most important events of the WOCE Data Products Committee meetings were special sessions (these were clearly not of interest to all DACs) to resolve DAC-to-DAC differences in expocodes, chief scientists, cruise dates, etc. In earlier WOCE days the most basic level of information had to be agreed upon; for example, DACs had their own cruise identifiers and a system common to all DACs had to be established. But for the WOCE Hydrographic Program cruises especially there was a great deal of cruise data tracked by DACs other than the WHPO (e.g., surface meteorology, ADCP, etc.). This inter-DAC reconciliation effort reached what to some might appear to be new levels of arcanity, but was absolutely essential in order to improve the joint utilization of WOCE data. In some cases it was the dogged persistence of just one or two persons (from the DIU) who saw that these issues were resolved. Although much of this business was accomplished electronically, the face-to-face discussions were extremely valuable. To accomplish this, it was required that the data specialists, not simply their scientific directors, meet. Hence, attendance at WOCE DPC meetings included a wide range of personnel, which strengthened discussion of other issues as well.

GLOBEC data management -- Dawn Ashby

GLOBEC Data Management

Dawn Ashby, GLOBEC IPO, November 2003

GLOBEC aims

The aim of GLOBEC is to advance our understanding of the structure and functioning of the global ocean ecosystem, its major subsystems, and its response to physical forcing so that a capability can be developed to forecast the responses of the marine ecosystem to global change.

GLOBEC has four primary objectives:

1. To better understand how multiscale physical environmental processes force large-scale changes in marine ecosystems
2. To determine the relationships between structure and dynamics in a variety of oceanic systems which typify significant components of the global ocean ecosystem, with emphasis on trophodynamic pathways, their variability and the role of nutrition quality in the food web.
3. To determine the impacts of global change on stock dynamics using coupled physical, biological and chemical models linked to appropriate observation systems and to develop the capability to predict future impacts.
4. To determine how changing marine ecosystems will affect the global earth system by identifying and quantifying feedback mechanisms.

GLOBEC structure

GLOBEC is organised into 4 research working groups, each examining a different aspect of GLOBEC science, plus 6 Regional Programmes focused on specific geographic regions and 18 countries with National GLOBEC programmes.

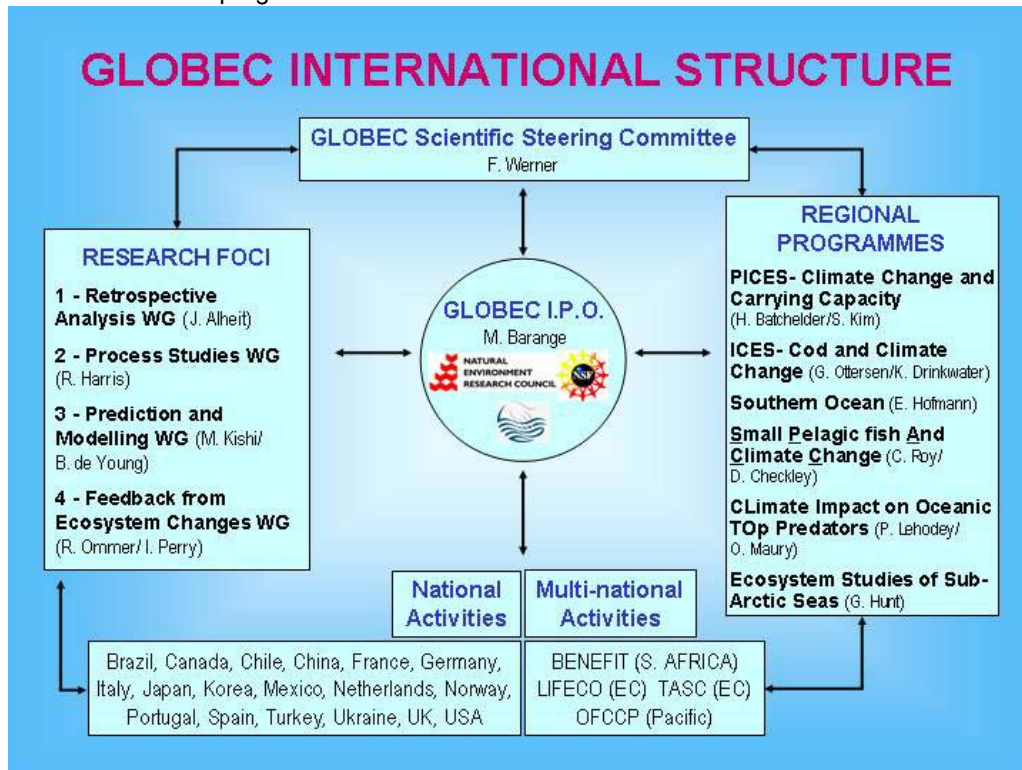


Figure 1. GLOBEC International structure

GLOBEC Data management

GLOBEC uses a decentralised data management system where metadata is held in a central database and individual projects are responsible for quality control and archiving their data. GLOBEC is a multidisciplinary programme which produces very varied types of data including hydrographic, biological, chemical, socio-economic and politically sensitive fisheries data, thus it is very difficult to incorporate the data into one database.

GLOBEC became one of the IGBP core projects in 1995 but it was not until 1999 that the IPO was formed. There were no formal data management activities for International GLOBEC before December 1999, when a data manager was appointed to work in the IPO. This was long after data collection had begun, and even after some of the GLOBEC projects had been completed, therefore it is difficult to obtain records of data for some of the early projects.

The GLOBEC Data Policy was drafted in early 2000; it focused on data and metadata sharing, the responsible archiving of data and inventory and cataloguing activities. It gives National and Regional GLOBEC programmes a framework from which to construct their own detailed data management policies and to address issues such as long-term archival of the data to ensure that GLOBEC makes a lasting contribution to marine science.

GLOBEC Metadata

GLOBEC metadata is stored in the Global Change Master Directory (GCMD) database. GLOBEC has its own portal into the database and scientists can either write their own metadata entries or submit information to the IPO. The GCMD uses a data format called a DIF (Directory Interchange Format), a standard used to create directory entries which describe a group of data. The DIF allows users of data to understand the contents of a data set and contains fields that are necessary for users to decide whether a particular data set would be useful for their needs. Six fields are required in the DIF; the others expand upon and clarify the information. Some of the fields are text fields, others require the use of valid values. The 6 required fields are:

1. Entry identifier
2. Entry Title
3. Parameters - include categories, topics, terms and variables. The parameters must be selected from the extensive list specified by the GCMD.
4. Data Centre
5. Summary
6. Document Author

Typology

In order to classify Internet data resources by geographic region a biogeographical typology produced by Daniel Pauly was used on the GLOBEC web pages as a cartoon on which to place useful links to data sources of relevance to the GLOBEC programme.

GLOBEC National, Multi-national and Regional Activities

The IPO maintains records of the activities of all GLOBEC programmes which was published as GLOBEC Special Contribution No.4, "Report on the Activities of the GLOBEC National, Multi-national and Regional Programme Activities" in 2001. The report was distributed to the entire GLOBEC mailing list of over 1500 individuals. Efforts are underway to update records of GLOBEC activities and an updated report will be published in early 2004.

Publications

At present there are over 600 publications in the GLOBEC database, which is an underestimate that we are trying to overcome through literature searches. Bibliographic details are held in an Endnote database and published on the GLOBEC webpages using Reference Web Poster. From the database it is possible to search for publications by authors, words in title or keywords and to download records directly into a Reference Manager or Procite database. Alerts have been set up with commercial databases to inform the GLOBEC IPO of new GLOBEC publications, and investigators can send details of their publications to the IPO via a form on the webpages.

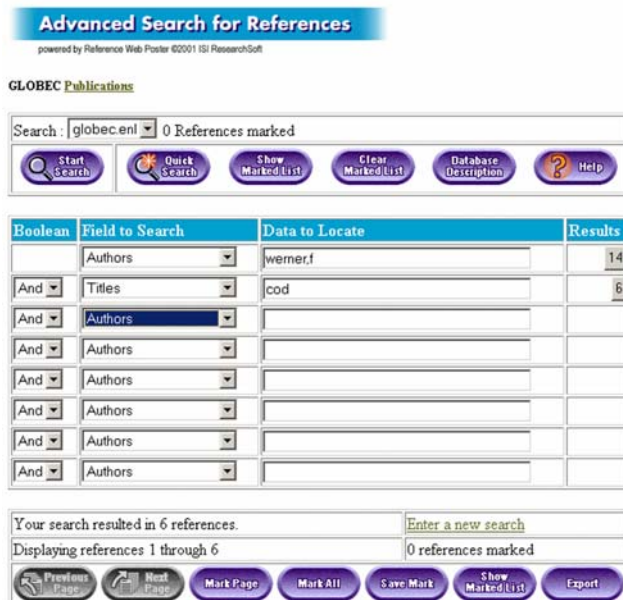


Figure 2. GLOBEC publication database

Responsibilities for data management

The GLOBEC Data Policy specifies who is responsible for each aspect of GLOBEC data management.

Individual scientists

Individual scientists have the responsibility for quality control of their data and making the data available within the timescale specified in the data policy for their programme.

National/Regional programmes

Each National and Regional programme should produce its own data management policy, either as a separate document or part of its implementation plan, to work alongside the GLOBEC Data Policy. This should address issues such as data sharing, archiving and cataloguing to ensure that the data are archived in a responsible manner.

IPO/Data Manager

The IPO/Data Manager is responsible for maintaining records of the activities of GLOBEC programmes and making these available to the GLOBEC community. This includes compiling details of each programme, encouraging and assisting scientists to write DIFs for the GLOBEC metadatabase and monitoring progress of data migration to permanent archives. The IPO/Data Manager ensure regular flow of information between the programmes and researchers, mainly via the newsletter and the GLOBEC Web pages.

Data Management Task Team

The Data Management Task Team (DMTT) was established in July 2000 to advise the IPO and Data Manager on suitable approaches for GLOBEC data management. The first meeting of the DMTT was held in October 2002 to coincide with the GLOBEC Open Science Meeting. As there is no specific funding for data management in the GLOBEC budget, it envisaged that the DMTT will continue to act in an advisory capacity primarily through e-mail contact, unless a specific need arises for a further meeting.

Scientific Steering Committee

The role of the GLOBEC SSC is to oversee GLOBEC data management activities and policies.

Achievements so far

GLOBEC metadata

The GLOBEC metadata portal at GCMD was set up in July 2000. At first investigators were slow to submit DIFs and most of the early records are due to the efforts of the GLOBEC Data Manager. National programmes are now starting to submit metadata to the database, in particular there has been a substantial input from US-GLOBEC. There are now 149 entries in the database, the majority being from the UK, USA and Regional Programmes.

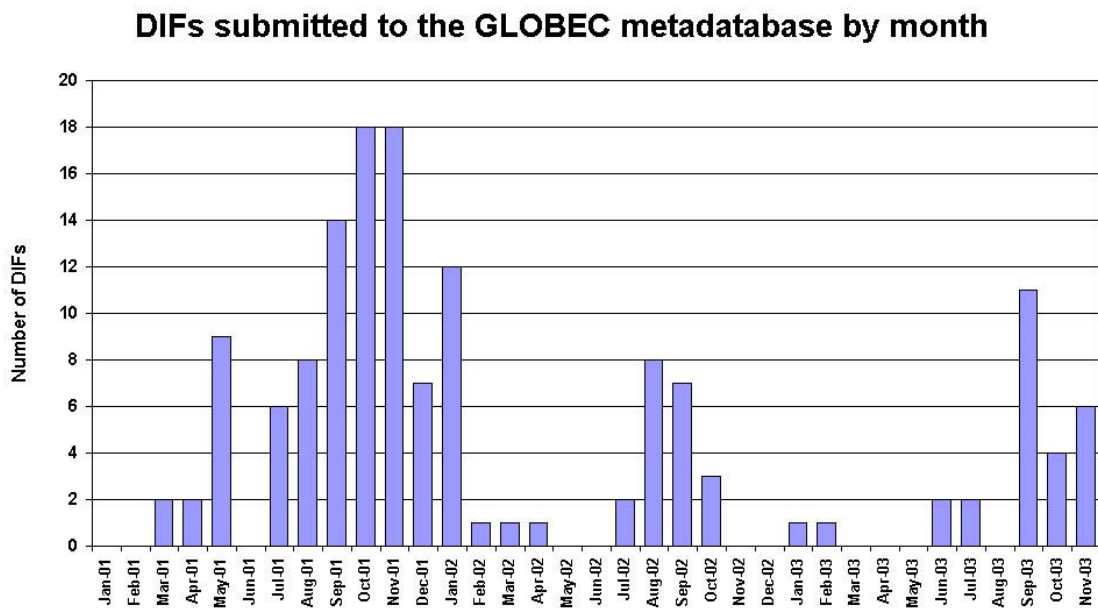


Figure 3a. DIFs submitted to the GLOBEC metadatabase by month

Number of DIFs in GLOBEC metadata database by country

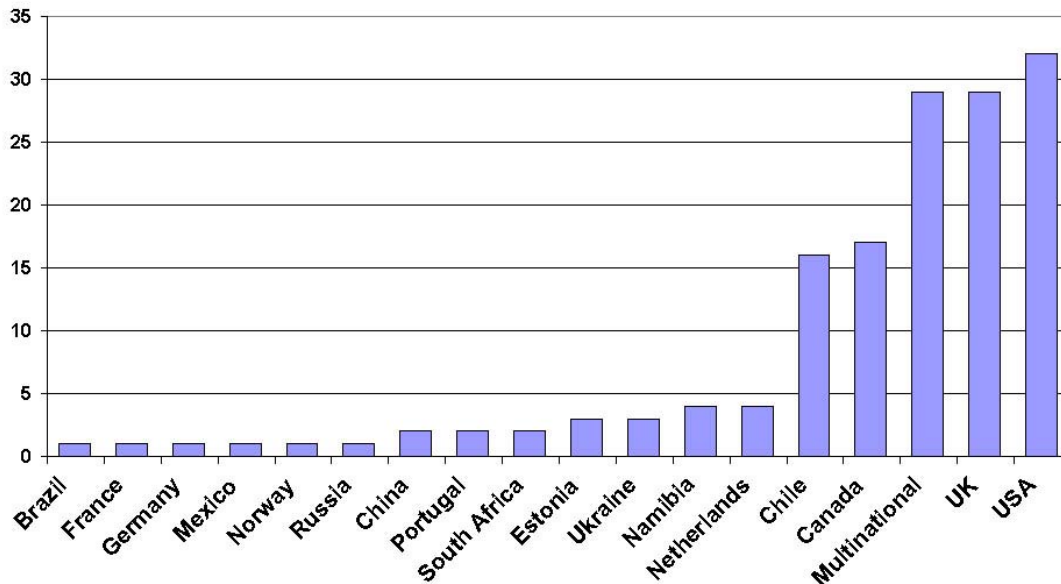


Figure 3b. DIFS submitted to the GLOBEC metadata database by country

Other products

A CD-ROM was produced in 2002 to coincide with the GLOBEC 2nd Open Science Meeting in Qingdao. The CD-ROM contains 3 GLOBEC products: an off-line copy of the GLOBEC website including all hyperlinked files, a copy of the GLOBEC publication list and a copy of the metadata file that were contained in the GLOBEC metadata portal. The objective of the CD was to facilitate access to GLOBEC products for those scientists who do not have reliable internet links and to encourage the community to continue supporting GLOBEC's data initiatives.

Lessons learned

It is important to have a data management policy established from the outset of the programme as it was found to be difficult to combine existing data management systems with the new metadata system. It has been difficult to obtain records of data produced by projects that were active before formal data management activities began.

Limitations to data sharing

The GLOBEC metadata portal provides a focal point where descriptions and location of GLOBEC data can be found; however, these records may not be complete for early projects. As the actual data is not held in a common format and the nature of the data is very diverse this may cause problems when sharing the data. Fisheries data may be politically sensitive, and countries may have reservations about sharing their data.

IMAGES data management -- Stefan Rothe

LOICZ data management -- Liana McManus

DATA MANAGEMENT IN LOICZ PHASE ONE

Dennis Swaney, Robert Buddemeier, Liana Talaue McManus, Stephen Smith, Hartwig Kremer and
Laura David

During phase one of the Land-Ocean Interactions in the Coastal Zone Project (1993-2002), two databases were developed: the nutrient budgets and global environmental typology databases. The attendant protocols in data management for both differ mainly because of differences in data sources. However, one database is linked with the other, and may be manipulated as an integrated database. The discussion below follows the guide questions provided by SCOR and details the protocols implemented by LOICZ. LOICZ hopes to improve and update these in keeping with the rolling-synthesis product-oriented approach it has adopted for phase two.

Basic features of data management

Currently, the LOICZ biogeochemical budgets program maintains a Website (<http://data.ecology.su.se/mnode>) with hierarchical, map-based linkages to individual Web pages describing each budgeted site. Each budget site webpage consists of a narrative overview and description of water, salt and nutrient budgets for the site. In addition, the basic data for all budget sites are tabulated in an Excel spreadsheet which is downloadable from the Website. A version of these data has also been entered into the LOICZ typology database, together with all of the other global data assembled by the typology group.

The georeferenced typology and budget data are searchable in a variety of ways, and are downloadable at: <http://www.kgs.ukans.edu/Hexacoral/Envirodata/envirodata.html>. The global environmental typology database is based on publicly available data. In its original form, it is not a project product. However, many aspects of the data processing and access system represent "value added" contributions to these data, and they are used to expand and interpret the biogeochemical budget data in ways that make them an integral part of the overall product. Data are available for download, along with some analytical tools, at the website indicated.

Quality control and data sharing

Biogeochemical budget site data required to construct budgets are organized by the individual contributors following LOICZ budget guidelines. Typically, a budget is constructed in a workshop format by scientists familiar with the estuarine system under consideration, and under the supervision of LOICZ project scientists. A short descriptive summary of the site is written, together with a brief analysis and interpretation of the nutrient budgets. These are reviewed by LOICZ scientists and revised, either during the workshop session or in succeeding weeks, and then posted on the budget Website. Occasionally, after the materials are posted, errors come to light as a result of an independent third party reading the materials. In such cases, either the budget authors or LOICZ personnel can be contacted to make the necessary corrections. When the budgets are corrected online, the corresponding changes are also made in the Excel summary spreadsheet.

Typology data sets are selected for appropriateness on the basis of consultation between the budget and typology teams, formatted into the half-degree grid cell format served by the Oracle database, and reviewed for coverage and consistency. Metadata offerings are based on the parent datasets. In addition to initial internal review, the project relies heavily on user feedback to identify problems with the data, whether these are errors, missing data, or user applications for which they may be appropriate.

Data submission to NODCs and WDCs

Because the nutrient budgets are based on previously existing data which are archived and authored by individual scientists, they are not submitted to any data center by LOICZ. In the case of the typology database, most of the typology data comes from, rather than goes into NODCs, WDCs or other providers. For both databases, the primary project output of integrated and at least partially interpreted results is not well suited for strictly data-oriented archival processes.

Budget for data management activities

Prior to 2003, one scientist (D. Swaney) was funded at approx. ¼ full-time equivalent to maintain the nutrient budget Web pages. For the typology, LOICZ IPO support ended at the end of 2002. Prior to that time, most of the ~\$20,000/year was devoted to data management in the larger sense, including development and maintenance of the user interfaces. Most of the major costs of Website and database support and development have been borne by collaborative funded projects (primarily “Hexacoral” with several others contributing), and by the host institution (Kansas Geological Survey). These total costs have been substantial, and currently are not met.

Involvement of Project SSC and IPO/Secretariat

The SSC and IPO essentially were uninvolved in data management issues, beyond guidance and consultation on presentation and products and support for the acquisition (workshop) process. Questions related to biogeochemical budget data format, quality, availability, etc., were dealt with by dialogue between Steve Smith, Fred Wulff (the PIs of the project) and Dennis Swaney (the manager of the Website/dataset). Typology data issues were dealt with by Bob Buddemeier in coordination with the budget team (above), Prof. Daphne Fautin (PI of the collaborative supporting NSF grants), and various technical support personnel.

Lessons learned

1. Data have errors (in the sense of mistakes), some of which can be caught and corrected before they enter the database, some of which after they enter the database, and some of which persist in the database. We attempted to minimize these via a fairly conventional review process, but it would be worthwhile to assess this process and consider additional measures for quality assurance. Cooperative QC by user and community expert feedback is extremely important, and needs to be facilitated so that data can be both promptly used and promptly reviewed.
2. Data have various levels of quality, which has several dimensions, including accuracy, precision, reliability, and data source. Initially, the project paid little attention to this question, rather choosing simply to accept or discard data or entire budgets on the basis of subjective criteria. As the project matured, a subjective ranking system was implemented in the budget dataset spreadsheet to attempt to grade the quality of various budgets. Future efforts at collecting and entering data should explicitly consider these dimensions.
3. The question of accuracy and precision of georeferencing and geospatial data is important, especially when working with datasets that have different inherent characteristics (point, polygon, areal) and scales (e.g., from 10s of m in position to half-degree in background characteristics). Advance planning for acquiring, evaluating and combining the geospatial components of various datasets is now seen as more important than originally recognized.

Limitations in data sharing and potential solutions

We make an effort to make the budget data available to interested users. The budget dataset spreadsheet is available to be downloaded from the Website. Two issues that could be considered as limitations come to mind:

1. Timeliness of updating the spreadsheet. As new budgets become available, they must be “vetted” and their data added to the spreadsheet. Depending upon the number of budgets “in the pipeline” and the time available, this can take time.

2. Appropriate use of the data by potential users. At least one “third party” has expressed interest in using the budget data for purposes of model validation or to develop statistical relationships between variables. The question has arisen as to how some of the values of variables are determined (i.e., are they measured, simulated, or are they themselves a result of some statistical or mathematical relationship, and thus not statistically independent of other data). Future efforts at collecting data should expressly consider the use to which such data can be put by third parties (and within LOICZ), and some indication of the data’s origin should be provided.

As in a number of user responses to the budget database, data analyses can be oriented towards identifying variables that can proxy some of the data requirements. Initially, there was a sense that the requirements of the nutrient budget model were simple enough. The number of missing empirical measures or if available were of low temporal quality encountered during the workshops provided the momentum and continuing interest to develop models for deriving some of the data inputs (e.g. runoff, anthropogenic nutrient inputs, among others)

Typology issues are similar, although the download and access process is much more advanced. Timely updating of the offerings as new datasets become available, while continuing to offer the previous datasets to serve those who may need to compare with previous results, is demanding but important. We see the optimal way to address this as joining into scientific consortia so that common facilities are supported by multiple users. This, however, is neither logistically nor financially trivial.

One of the limitations in both domains (budget and typology) relates to the geospatial linking and identification of the various datasets and points. We find that GIS techniques offer vital analytical, selection, dissemination, and visualization capabilities, and feel that standardizing and expanding the use of these approaches are critical.

Potential directions in developing data products

Rates of change in hindcast population and land use modeled data is a potential data product that can be provided using historic terrestrial climatologies from Willmott et al. Currently, these are served as means and averages in the typology database. Interactive mapping through display and download interfaces may be done should resources for both development and maintenance be made available.

In terms of increasing spatial resolution, this can be done with some nested datasets, which hopefully will include refined data on coastal complexity and oceanography. LOICZ has offers to integrate into the current holdings a number of regional datasets (e.g., DINAS Coast and New Zealand estuaries). With the latter, data users can experiment across spatial scales to see whether modeled patterns are robust or change with spatial resolution.

Linking biological features with the physical and chemical signatures in coastal areas is a wide open area for interpreted data development. Mapping out fisheries variables, or indicators of population change of indicator organisms in relation to coastal geochemical changes are potential projects that would require strong partnerships among interested global change programs. Given what LOICZ has initiated, these potential products are realizable in the medium term with an appropriate investment in expertise and resources.

Finally, elucidating the influence of human activities in altering rates of coastal change including its biology and spatial boundaries, may be possible with the integration of appropriate socio-economic variables in the databases. However, the scales at which these are measured (e.g., national scale) are often too coarse to relate with environmental variables that are available at watershed or coastal basin scale. New protocols of gathering and managing socioeconomic data need to be developed to allow for seamless data integration and appropriate data assimilation and interpretation.

The World Data Centre network-- Ferris Webster

World Data Centers relevant to marine science

Introduction

The World Data Centers (WDCs) operate with the guidance of a panel under the auspices of the International Council for Science (ICSU). The centers were created nearly fifty years ago, after the International Geophysical Year, to deal with geophysical and solar data. Their support comes from the host nations.

The WDCs originally centered on physical sciences, such as solar-terrestrial physics, geomagnetism, oceanography, and meteorology. Today their range in disciplines is more broadly environmental, with 52 centers in 12 countries.

The network is growing, expanding to meet new needs. The growth has been stimulated by the desire to meet the needs of new ICSU programs, particularly those in IGBP and WCRP. Some examples of new centers created during the last decade are WDCs for biodiversity, climate, human interactions in the environment, land processes, paleoclimatology, and soils.

Though the WDC System has evolved to meet new needs, there are still challenges. For example, nearly all the centers are in developed countries in the Northern Hemisphere. To improve its value to scientists in developing countries, the System is establishing partnerships involving an established data center and a "mirror site" in a developing country.

WDCs also are evolving to take advantage of new technology. The internet has become the principal means for distributing data, though scientists in some regions do not yet have the capability to use the internet. The need to create an effective WDC Web portal remains a significant technical challenge for the System.

Relevance to marine science

More than 15 WDCs have marine holdings:

- Atmospheric Trace Gases, Oak Ridge [marine CO₂ data]
- Biodiversity & terrestrial ecology, Denver [fisheries and aquatic resources including some marine]
- Climate, Hamburg [ocean models, IGPP coupled models, climate-change data]
- Glaciology, Boulder, Lanzhou, Cambridge [sea ice]
- Marine Environmental Science, Bremen & Bremerhaven (WDC-MARE) [environmental oceanography, marine geology, paleoceanography, and marine biology]
- Marine Geology & Geophysics, Boulder, Gelendzhik [marine geology and marine bathymetry]
- Meteorology: Asheville, Beijing, Obninsk [air-sea interaction data, TOGA datasets, COADS]
- Oceanography: Obninsk, Silver Spring, Tianjin [all types of marine data]
- Paleoclimatology, Boulder, Nairobi [marine sediments, corals, sea level]
- *Satellite Information, Greenbelt [remotely-sensed marine data]*

The WDC system is integrated with the data centers of the International Data and Information Exchange (IODE) program of the Intergovernmental Oceanographic Commission (IOC)

IGBP-WDC partnerships

The WDCs want to work with ICSU programs. For the collaboration to be successful, there should be a real partnership between the IGBP project and the WDC. The WDCs should be more than simply a place to dump the data after a research program is done. Project scientists should work with the data center in the planning stages of the program. As the program is underway, the data center should be

a partner in developing a system to assure effective access to the program results. Some examples of successful partnerships are:

- TOGA and TOGA/COARE datasets & metadata are held by the WDC for Meteorology, Asheville
- WOCE datasets are held by the WDC for Oceanography, Silver Spring. The WOCE Data Information Unit continues to operate on-line from the WDC.
- JGOFS datasets will be held by the WDC for Marine Environmental Science, Bremen, and the WDC for Oceanography, Silver Spring.

Data access

The WDC principles on data access provide assurance of access to all. In part the principles read: "WDCs will provide data to scientists in any country free of charge or at a cost not to exceed the cost of copying and sending the requested data."

"WDCs operate ... for the benefit of the international scientific community and provide a mechanism for international exchange of data in all disciplines related to the Earth, its environment, and the Sun."

Principles at risk

Intellectual property rights are jeopardizing the ability of the WDCs to freely obtain and distribute data. This situation is in conflict with ICSU's policy, which recommends full and open access to data for scientific purposes. New legislation, both national and international, may subject datasets in WDCs to new intellectual property restrictions. For example, under some proposed new laws, users would have to pay and/or obtain permission to use more than an insubstantial amount of a database.

The threat to access to scientific data is real: the European Community enacted a Database Directive in 1998. A treaty proposed at the World Intellectual Property Organization in 1996 was withdrawn, partly due to pressure from the research community. The US Congress has been considering new database legislation for several years, though so far no legislation has emerged.

The outlook for the future

The opportunities and the need for the World Data Centers has never been greater. The World Wide Web is now used as the primary means for dissemination of data. New global science programs provide the occasion for the WDCs to continue to diversify their coverage.

Though technology offers many options for data sharing, the centers are still needed. The staffs at the centers provide expertise with the holdings. The centers assure long-term preservation and guaranteed access to datasets. The WDC system assures international access under the ICSU policy on data access, and assures availability of data to future generations of scientists in a changing world.

Ferris Webster, 10 December 2003

Annex: Principles and Responsibilities of ICSU World Data Centers

The basic principles and responsibilities of the international exchange of solar, geophysical and environmental data through the World Data Centers have carried forward under ICSU rules, essentially unchanged since the establishment of the WDC system for the IGY.

1. World Data Centers are operated for the benefit of the international scientific community. They are supported by national organizations according to these Principles laid down by the ICSU Panel on World Data Centres.
2. The resources required to operate WDCs are the responsibility of the host country or institution, which is expected to provide these resources on a long-term basis. If for any reason a WDC is closed, the data holdings shall be transferred to another WDC.
3. WDCs will, subject to their financial resources, accept data according to the data management plans of appropriate ICSU scientific programs or monitoring activities, and store these data safely and in good condition. WDCs may enhance their holdings by seeking and collecting related data sets. They may prepare higher-order data products such as indices of activity and collated or condensed data sets.
4. WDCs will prepare and publish catalogs of their data holdings, or otherwise make freely available information on their holdings, e.g., by electronic access.
5. WDCs will exchange data among themselves, as mutually agreed and whenever possible without charge, to facilitate data availability, to provide back-up copies, and to aid the preparation of higher-order data products.
6. No confidential or security-classified data are to be held in a WDC.
7. Data may be subject to privileged use by their originators, for a period to be agreed beforehand, and not to exceed two years from the date of acquisition by the WDC.
8. WDCs will provide data to scientists in any country free of charge, on an exchange basis or at a cost not to exceed the cost of copying and sending the requested data. Additional charges may be made for special services, or for acquiring data from outside the WDC system.
9. WDCs will accept any scientist as a visitor to work on site with data holdings held under WDC auspices.
10. WDCs will report to the ICSU Panel as requested.

The IOC IODE network -- Lesley Rickards

Data Management in GOOS/JCOMM -- Lesley Rickards

Ocean Biogeographic Information System (OBIS) -- Geoff Boxshall

Ocean Biogeographic Information System (OBIS) - www.iobis.org

The Ocean Biogeographic Information System (OBIS), the information component of the Census of Marine Life (CoML), is a rapidly developing international science program to provide access to data content, information infrastructure, and informatics tools, (maps, visualizations, and products from models) through an on-line, dynamic, global 4-D (the three dimensions of space plus time) atlas of biogeographic information. The atlas will be used to reveal spatial/temporal patterns, generate new hypotheses about the global marine ecosystem, and guide future field expeditions. The on-line, digital atlas developed by OBIS is expected to provide a fundamental basis for societal and governmental decisions on how to sustainably harvest and conserve marine life. The scope of OBIS offers new challenges in data management, scientific cooperation and organization, and innovative approaches to data analysis.

OBIS is an Associate Member of the Global Biodiversity Information Facility (GBIF). OBIS, under the direction of an International Committee and Secretariat, has been rigorously pursuing research and development in the following directions:

Data access: gathering comprehensive, accurate, quality-controlled digital data

Species data are needed to study food webs, population dynamics, evolutionary history, habitat, biogeography, species introductions, and criteria for establishment of marine reserves. Even in the case of broad-scale questions, a “black box” approach to taxonomy is inadequate because particular species are likely to play roles disproportionate to their abundance. Indeed, our ability to address critical ecological questions hinges on our ability to know what organisms we are dealing with. Historic development has brought about multiple views of species definitions, which further complicate the issue of correctly using data from existing collections. Adequate quality control is achieved through direct involvement of taxonomic authorities for each group. OBIS has mobilized the marine systematic community to digitize and store geo-referenced distribution data on accurately-identified species. The OBIS community is also providing expertise on marine systematics to Species 2000, ITIS (the Integrative Taxonomic Information System) and the Electronic Catalog of Names of Known Organisms Program of GBIF.

Data integration: integrating heterogeneous data sources

With the advance of new technologies, digitization of existing records, and active field explorations such as those in the CoML, the production of heterogeneous data with complex interrelationships will increase daily. The volume and diversity of these data constitute an urgent research issue in data integration and interoperability. The major topics include: 1) community-endorsed global data and metadata standards, 2) new tools and algorithms for semantic mapping and integration, 3) efficient algorithms for data aggregation, and specifically, geospatial and temporal data aggregation. All of these topics are central in building the infrastructure for global biodiversity and environmental information systems.

Data analysis: providing the ability to discover scientifically important patterns and unique events

The emergence of GOOS and GBIF indicates a paradigm shift in earth system sciences. The ever-increasing volume of ecosystem data and their successful integration pose new challenges to researchers. New, scalable algorithms and tools must be developed to efficiently search for scientifically interesting, temporal-spatial patterns and to identify unique, sometimes disruptive, ecosystem events in large, integrated databases. Data mining techniques are inductive in nature and the main purpose of scientific data mining is to assist scientists at the initial stage of scientific discovery, i.e., generating hypotheses based on observations and heuristic relationships. Patterns identified with automatic mining techniques have to be examined carefully by domain experts and further validated by deduction-based methods. We need to combine data mining, traditional statistical analysis, and mathematical modeling to understand complex marine ecosystems and to formulate reasonable predictions. OBIS promotes a synthetic and cooperative approach to ecosystem study and will serve as a global forum for integrated ocean biodiversity study and biogeographical research.

Data visualization: developing a new generation of marine Geographical Information System (GIS) and other visualization tools

Data visualization is an important part of the knowledge discovery process. It is particularly important in marine ecosystem studies because of the spatial-temporal nature of ocean ecosystem data. Current GIS systems cannot deal well with 4-D data so new data structures and algorithms must be developed. Meanwhile, many existing GIS tools cannot meet the user demand for Internet-based mapping services. OBIS is actively working in these areas and some products are already available on the OBIS portal.

Data policy: summarized as free and open data

OBIS is conceived as a federation of distributed data holders. OBIS data are open access to the public all over the world. OBIS is an Associate Member of GBIF and subscribes to the GBIF policy on data sharing (<http://www.gbif.org/moufree/mou2.htm#§8>). The data provider retains all rights to the use of the data they provide. By publishing their data on-line through the OBIS portal, they allow others the right to analyse and integrate their data with other data, both biological and environmental.

SOLAS Data Management -- Doug Wallace

SOLAS: Data and Model Management

Background and Needs

The implementation of SOLAS will involve the collection of large quantities of environmental data under separate nationally and internationally organised projects. These data will be collected from process studies and experiments, time-series studies and large-scale surveys. Similarly, SOLAS will make use of a hierarchy of different modelling approaches (see section x.xx. In most cases, the utility of the models and data involved in these projects will extend beyond the projects themselves and be of interest to other investigators. Further, many SOLAS data will be more useful when combined with, or compared against, other data and models including non-SOLAS data. Scientific findings and conclusions derived from SOLAS projects should be available for assessment by independent scientists: this implies that the underlying data and/or models must be readily accessible.

Increasingly, key management issues are based on model results as well as data. Model-derived results are also used extensively as input for other models, for scientific planning and for policy decisions. SOLAS must therefore ensure that the models developed and/or used in SOLAS as well as the data collected are documented sufficiently to allow independent evaluation and that models and/or model results as well as data be readily accessible to the scientific community for assessment and interpretative purposes.

- Data and model management are therefore critical logistical tasks for SOLAS.

SOLAS science will involve collection of complex and sophisticated data sets. This will include difficult, error-prone, measurement of biological, physical and chemical parameters. Because SOLAS is an international, multi-investigator program, such measurements will be made all over the world, at different times, by different groups, often using different equipment and techniques.

- Attention to data quality management will be critical for the scientific integrity and success of SOLAS.

Data Management Principles:

Certain basic principles should guide the development of data management procedures in SOLAS:

Do not reinvent the wheel.

The past decade of global change research has driven the evolution of a wide range of data management approaches, policies and data centres. This evolution process was should not be repeated unnecessarily: SOLAS should make use of the lessons learned by, for example, WOCE and JGOFS. SOLAS should also avoid developing its own, independent approaches except where absolutely necessary. Rather SOLAS should, to the maximum extent possible, take advantage of existing data management procedures, methods, data centres, etc.. This includes adaptation and extension of such approaches to SOLAS needs, as necessary. For a wide variety of SOLAS data types and data streams, the data management methods developed within the World Climate Research Program may be applicable. A critical review of WCRP data management procedures with respect to their suitability for SOLAS purposes should be conducted.

Plan ahead for rapid data assembly.

Data release, assembly and sharing is often an afterthought, considered only upon completion of a project. SOLAS must alter this recurrent pattern. Specifically, any SOLAS project must consider the problems of data management and assembly in detail at the project design and proposal stage. This must go well beyond simply constructing a timeline and making promises as to when data will be released. Rather each proposed SOLAS project should consider explicitly the common problems associated with data management, and how these will be addressed. The description of data

management procedures in proposals should ideally be at the same level of detail and thought as the description of experimental and measurement protocols. SOLAS proposals should therefore be very explicit and detailed as to:

- What types of data are going to be collected?
- What data structures are appropriate for these data?
- How much data of which types will be collected?
- How will data from the same project be identified, organized and merged?
- How will data be quality-controlled?
- What is the timeline for preparation of 'final' data?
- What might go wrong and delay the preparation of final project data?
- What metadata will be provided and when?

If all SOLAS project proposals were specific on these issues, then the PIs would have a much better realization of what is expected of them and could plan and budget accordingly.

Data managers should support data gatherers.

Most data centers consider their primary task to be the 'archiving' and dissemination of data and metadata. One often hears frustration from data center personnel and professionals concerning the pace of data reporting from PIs. The bottlenecks in data reporting and problems of data availability are generally attributed to data originators. Frequently this is attributed to reluctance of the PI to share data and this is, indeed, frequently a problem. However in other cases, delays are associated with a lack of data management skills and support available to the PI. Quite simply: some PIs are not very good with data management and have no assistance in this area. Data centers on the other hand, have considerable expertise and capability in this area.

Within SOLAS, data centres should be encouraged to assist PIs with the development of tools and procedures that will make data reporting tasks easier and more efficient. They should offer to work cooperatively with the PIs to help make data available, rather than simply wait for it to be delivered. For example, data centers could offer a program of data manager and /or programmer support to advise and assist PIs directly with data management procedures. Such a proactive approach would have the added benefit of making data centres more familiar with the types of data they are handling.

SOLAS should reward excellence in data collection and data release.

SOLAS, perhaps more than any other global change research program, will depend on excellence and innovation in data collection. Many of the parameters to be measured will depend on the development and application of innovative technologies (Focus 1 and 2). In other cases, attention to data quality as well as data collection efficiency will be required to advance SOLAS science (e.g. Focus 3). SOLAS should ensure that such efforts are encouraged, and that success in these areas is rewarded.

Part of the data reporting problem in global change science may be that data reports and the documentation of experimental findings are commonly published as 'grey literature'. Data reports and data documentation are therefore considered to be of lesser importance in comparison to the 'peer-reviewed literature'. This is unfair to those who choose to collect data carefully and document it thoroughly, is inconsistent with the widespread view that 'metadata' and rapid data release is critical for scientific progress. The lack of peer-review of data also means that there is generally no formal assessment of data quality.

It is illogical or at least inconsistent to classify data reports and metadata as being 'grey literature' and, at the same time, argue that data release, data documentation and data quality are 'fundamental' and 'essential' for the progress of science. SOLAS should tackle this inconsistency head-on and creatively. In many other areas of science, this downgrading of data collection and observation description does not occur. For example, papers detailing routine measurements of various thermodynamic properties (e.g. gas solubilities) are regularly published in the peer-reviewed literature. So are descriptions of newly discovered species. Why should well-documented observations and findings from oceanographic process studies or time-series not be similarly published, reviewed and cited?

There is at least one potentially simple, but untested mechanism to encourage data submissions. SOLAS could establish, in cooperation with a data center or a scientific society, a peer-reviewed electronic journal that is dedicated to the publication of project data sets and metadata. These data sets and metadata should, themselves, be peer-reviewed in order to ensure quality and adequate documentation. This peer-review will help to ensure data quality and adequate documentation of SOLAS data (see below). In turn the contributing investigators will have a citable, peer-reviewed publication as reward for their (considerable) efforts. The establishment of a 'Journal of SOLAS Data' would therefore be beneficial, not only in promoting and rewarding the release of key data sets (the data originators' contributions will now be peer-reviewed publications), but would for the promotion of data quality and data documentation. Admittedly unclear with such an approach is what would happen to data sets that fail peer-review or that are not submitted via this route. Presumably, a parallel 'grey literature' pathway to data submission/dissemination must also be maintained.

Specific Steps Towards Data Management:

SOLAS should:

Establish an international SOLAS data management task team with at least one paid staff member.

This team should:

- 1) Evaluate and document the likely data needs of SOLAS (data types, data quantities, data sources, metadata requirements)
- 2) Conduct a review and intense discussion with the WCRP and JGOFS data management community concerning lessons learned and present data management policies, problems and solutions. Identify the potential for SOLAS data management within existing structures and programs.
- 3) Evaluate the feasibility and benefits of establishing a peer-reviewed 'Journal of SOLAS data'.
- 4) Write a detailed SOLAS data management policy. This policy to include time-limits, enforcement procedures, access rights, metadata requirements, reporting requirements and procedures. Potentially, this document to include guides to data organisation, metadata requirements, etc.
- 5) Initiate negotiations with data centers re: possibilities for their direct support of individual PIs with respect to data management procedures and tools.
- 6) Host a workshop for the modelling community in order to develop and write a practical model management and documentation plan.

Data Quality Management

In addition to establishing clear guidelines for data management, SOLAS must also establish procedures for assessing and controlling data quality. Once again, lessons learned during WOCE and JGOFS can be used to address such issues.

Data quality management should be addressed by:

- 1) establishing a SOLAS data quality task team.
- 2) establishing clear quality targets for SOLAS data.
- 3) documenting recommended protocols for 'standard' SOLAS measurements.
- 4) providing support for technical workshops, training sessions and calibration and intercalibration activities.

IMBER Data Management Julie Hall, Chair, IMBER Transition Team

The Integrated Marine Biogeochemistry and Ecosystem Research (IMBER) project is a new activity under development by SCOR and IGBP. The goal of IMBER is to understand how interactions between marine biogeochemical cycles and ecosystems respond to and force global change. IMBER's themes are

- Interactions between marine biogeochemical cycles and ecosystems
- Sensitivity of biogeochemical cycles, ecosystems, and their interaction, to global change
- The role of the marine biogeochemistry and ecosystems in regulating climate
- Interactions between marine biogeochemical cycles, ecosystems and the human system.

It is obvious from the breadth of these themes that IMBER will need to manage a large variety of data types, including the following:

- Field data from cruises, moorings, drifters
 - Physical
 - Biological
 - Chemical
- Laboratory and mesocosm experimental data
- Remote sensing data
- Socioeconomic data
- Model-derived data

For each of these data types, IMBER will encourage of both collection of new data and “data mining” and “data rescue” for existing data that are not widely available for analysis and use in models. The IMBER Transition Team, and SCOR and IGBP staff have taken several steps to help IMBER develop an appropriate data management plan and structures:

1. Meeting at the OCEANS Open Science Conference—Discussions of an IMBER data management plan began at the OCEANS Open Science Conference in January 2003. A small group of individuals interested in and/or involved in data management met briefly to identify the most important issues. Two meeting attendees offered to draft a data management plan for the OCEANS project.
2. Preparation of draft document—Bernard Avril and Nicolas Dittert prepared the draft document and forwarded it to the OCEANS Transition Team (TT). This document was edited by Ed Urban and presented to an OCEANS editorial meeting in Washington, D.C. in May 2003.
3. Banff workshop—IGBP scheduled a half-day workshop on data management at the IGBP Congress in Banff, Canada, in June 2003 that resulted in a two-page document of findings and recommendations. The OCEANS TT was represented at this meeting (OCEANS became IMBER at the IGBP Congress.)
4. The Banff workshop recommendations were included in the draft IMBER *Science Plan/Implementation Strategy* (SP/IS) that was posted on the Web for comments by the international ocean sciences community.
5. The SCOR/IGBP data management coordination meeting was held in Liverpool, U.K. in December 2003 and resulted in specific recommendations for project data management template for transmittal to the projects for consideration.
6. The IMBER SP/IS was in the final stages of drafting at the time of the data management coordination meeting, so the principles agreed by meeting participants are included in the plan that was sent to review.
7. Develop a detailed Data Management Plan for IMBER. Further development of IMBER data management will depend on additional discussion by the IMBER Scientific Steering Committee and Data Management Committee.

CLIVAR Data Management -- Howard Cattle

CLIVAR Data Management
Howard Cattle and Katy Hill
International CLIVAR Project Office
Southampton Oceanography Centre, UK

CLIVAR's data management policy is articulated in its Initial Implementation Plan which itself gives broad guidelines on CLIVAR data management. A CLIVAR Data Task Team was set up to:

1. define CLIVAR's requirements for a data and information system
2. assess the extent to which existing data management systems meet the CLIVAR requirements
3. ensure the rapid, responsive delivery of data and data products
4. ensure the secure but accessible archival of CLIVAR data
5. ensure the delivery of information on the location and availability of CLIVAR data
6. make recommendations on actions that need to be taken to ensure an adequate CLIVAR data and information system.

This group had one meeting and was then disbanded. The key recommendations of the panel have since been implemented by the International CLIVAR Project Office (ICPO). These were:

- to establish data/products liaison members for each of CLIVAR's panels and working groups with a remit to identify key data and data products relevant to CLIVAR
- given CLIVAR's key role in the role of the oceans in climate, to transition the existing WOCE Data Assembly Centres (see below) into CLIVAR Data Assembly Centres.

The ICPO is now working, through an identified member of staff with responsibilities for liaison over CLIVAR data issues, to develop the activities of the Data Assembly Centres for CLIVAR and proposals for a Data Assembly Centres Workshop are currently being developed. CLIVAR is also working with IOC on development of a joint hydrographic and carbon database linked to the WOCE repeat sections. Because of the broad nature of CLIVAR, the current strategy is for CLIVAR to work through a distributed network of data centres accessed via links on CLIVAR Data Information Pages of the CLIVAR web site. These pages have been established and are under active further development in consultation with the CLIVAR panel and working group liaison members. By its very nature, many of the data sets CLIVAR deals with are "CLIVAR-relevant" rather than produced by the CLIVAR project per se. A current exception is the data produced under the CLIVAR Variability of the American Monsoon (VAMOS) Panel field programmes which are managed and archived through a VAMOS Project Office run by UCAR's Joint Office for Science Support (JOSS).

At its last meeting, the CLIVAR Scientific Steering Group agreed that data management in CLIVAR would be driven by a new panel, the CLIVAR Global Synthesis and Observations Panel. The Scientific Steering Group also agreed on the need for a focussed CLIVAR data management workshop and for the development of a coordinated plan drawn together through the help of a data consultant. These issues are being taken forward currently through the ICPO.

GEOHAB data management -- Wolfgang Fennel

(Draft report based on the GEOHAB SP and IP)

Introduction

The overall scientific goal of GEOHAB is to: **Improve prediction of HABs by determining the ecological and oceanographic mechanisms underlying their population dynamics, integrating biological, chemical, and physical studies supported by enhanced observation and modelling techniques.**

The programme elements of GEOHAB and their overarching questions are:

- **Biodiversity and Biogeography.** *What are the factors that determine the changing distribution of HAB species, their genetic variability, and the biodiversity of associated communities?*
- **Nutrients and Eutrophication.** *To what extent does increased eutrophication influence the occurrence of HABs and their harmful effects?*
- **Adaptive Strategies.** *What are the unique adaptations of HAB species and how do they help to explain their proliferation or harmful effects?*
- **Comparative Ecosystems.** *To what extent do HAB species, their population dynamics, and community interactions respond similarly under comparable ecosystem types?*
- **Observation, Modelling, and Prediction.** *How can we improve the detection and prediction of HABs by developing capabilities in observation and modelling?*

Example ecosystem types as defined by their bathymetry, hydrography, nutrient status, productivity, and trophic structure, include the following:

- Upwelling systems, such as those off the coast of Portugal and Spain, Peru, Mazatlan in Mexico, the west coast of the United States, Australia, Japan, West Africa, and Southern Africa.
- Estuaries, fjords, and coastal embayment systems, as in the United States, Canada, Australia, southeast Asia, Philippines, Mexico, Scandinavia, and Chile.
- Thin-layer producing stratified systems occur along most coasts, including the Atlantic coast of France, Sweden, California, and in East Sound, Washington.
- Coastal lagoon systems such as in the United States, Mexico, Brazil, and France.
- Shelf systems affected by basin-wide oceanic gyres and coastal alongshore currents such as off the northwestern European coast, the Gulf of Mexico and Gulf of Maine in the United States, and off the coast of southeastern India.
- Systems strongly influenced by eutrophication, such as in Hong Kong, Black Sea, Baltic Sea, Adriatic Sea, Seto Inland Sea of Japan, and the mid-Atlantic regions of the United States.
- Brackish or hypersaline water systems such as the Baltic Sea, St. Lawrence, Dead Sea, and Salton Sea.
- Benthic systems associated with ciguatera in the tropics or DSP in temperate waters.

The broad range of systems and the multidisciplinary approach of GEOHAB projects indicate the specific requirements for the development of data management plans. Some problems can be studied on the basis of single **species of interest** embedded in a physical chemical and biological environment. Other problems require consideration of parts of the **food web**, including **species-species and benthic-pelagic interactions**, and may involve **complex life cycles**.

Data Types in GEOHAB Projects

Data types are diverse and to some degree site- and project specific. Data from field studies ranging from physical and chemical quantities to biological parameters that characterize populations or parts of the food web are measured in different sites/ systems and need an appropriate management.

Genetics and Biogeography

Data on

- the genetic variability of HAB species and their biogeography
- cellular properties of HAB species: morphology, molecular genetics, biochemical composition, toxicity
- HAB taxa at various taxonomic levels, population genetics of HAB species over the annual cycle
- global biogeographics of key HAB taxa

(biogeographical, taxonomic and phylogenetic relationships among HAB taxa, the data are often presented as cluster diagrams)

Data on

- distributional pattern of HAB species (water column, benthic habitats, and sedimentary record)
- life cycles, growth and mortality
- biogeographical range of HAB species, (migration by currents, ships)

Eutrophication and HABs

Data on

- toxin production of HAB species at different stages and nutritional conditions
- the responses of the harmful properties of HABs to varying nutrient inputs and in relation to competing organisms
- noxious foams and scums

Regional oceanographic regimes

Data of physical processes (oceanographic regimes of ecosystems) and their effects on HAB
Gridded data of chemical biological variables at spatial and temporal scales consistent with those of the physical variables

Types of measurements:

- Chemical data (stations, moorings, underway measurements)
- Biological data (stations, remote sensing, tank experiments, labs)
- Physical data (stations, moorings, underway measurements, remote sensing)
- Modelling (gridded model data of all state variables and physical quantities, tables of parameters sets for models, animated pictures)

Data Management

It is stated in the *GEOHAB Implementation Plan* that the development of an appropriate GEOHAB data management plan is a fundamental and critical activity upon which the ultimate success of GEOHAB will depend.

GEOHAB data are relevant to scientists and managers beyond the GEOHAB community. Therefore, GEOHAB will participate in co-operative non-governmental and intergovernmental data management systems. GEOHAB will co-operate with the framework for research data being developed for SCOR and IGBP projects.

GEOHAB will also participate in data management processes of the International Oceanographic Data and Information Exchange (IODE) activity of IOC.

The Intergovernmental Panel on Harmful Algal Blooms (IPHAB) has recommended that “the IOC ensure that data-quality management and data exchange relevant to GEOHAB be given due consideration, in accordance with the Terms of Reference for the Group of Experts on Biological and Chemical Data Management Exchange Programme (GE-BCDMEP), and that a GEOHAB representative be included in the GE-BCDMEP.”

Each focused GEOHAB Open Science Meeting will be asked to discuss data management and to include data management plans within the research plans produced. Each GEOHAB project should address the long-term archival of observational data and data products to ensure a lasting contribution to marine science.

GEOHAB will use a decentralised data management and distribution system with a centralised index. The components, centralised under the supervision of the IPO, will include a comprehensive inventory of databases relevant to GEOHAB, as well as meta-data, with links to their locations and contact persons. All investigators should be prepared to share their data and data products within two years from the time those data are processed, and should recognise the “proprietorship” (rights to first publication or authorship) of data acquired from other investigators.

Data management issues will be handled by a small GEOHAB Data Management Committee, which will be responsible for ensuring that the GEOHAB data management policy is followed by participating projects and will assist the International Programme Office in data-related issues. The GEOHAB data management policy will be posted on the GEOHAB Web site.

Identification of Protocols and Quality Control

GEOHAB encourages the use of existing standard protocols and guidelines for sampling and experimental methods. GEOHAB investigators retain the primary responsibility for quality control and assurance. It is essential that the methods adopted to ensure quality control and the protocols used for data collection are fully documented in information files (meta-data) accompanying data sets.

Open science meetings for the Core Research Projects will be asked to specify core parameters that will be measured initially in each location, as well as standard measurement protocols. GEOHAB recommendations on methods and measurements will be disseminated through the IPO and GEOHAB Web site. Task Teams will be established, when necessary, to define methods to be applied or recommended in GEOHAB projects and to organise inter-comparison of methods and models.

GEOHAB modelling activities depend on data sets for initialisation, forcing and validation of models. Model output, in particular gridded data sets, should be provided in generic formats, for example, NetCDF, and support visualization by animated data products. GEOHAB will identify needs for development models and relevant existing modelling activities through a Task Team that has the responsibility to organise model inter-comparison exercises, including comparison of predictive models for HABs.

The Indian NODC approach to data management -- Jaswant Sarupria

Report on oceanographic data/information management activities at Indian Oceanographic Data Centre (IODC)

1. Brief History:

IODC was established in 1964 at NIO, Goa, India. NIO is a premier institute for oceanographic research and development. It has highly qualified staff of scientific (195Nos), and technical (236nos). NIO published on average more than 100 research papers, articles, and technical reports on oceanography annually. IODC is having four scientific qualified staffs and four technical & supporting staffs. IODC is using computer systems DEC ALPHA with 64-True Unix system and PCS AVIIION 3600 Unix system with open ingress RDBMS. Also using Oracle-9i on true Unix system and Oracle Internet Developed Suit (IDS-9i) on windows platform.

2. IODC Activities:

Processed information/ data are utilized to satisfy the oceanographic users society as per their requirements under the ocean information services. We have supplied oceanographic data and data products to more than 100 users per year. These agencies are working in the field of research and development, educational, defense, industrial and data management sectors. We have developed and updated thirteen data bases for thirty six oceanographic parameters for the water column in Arabian Sea, Bay of Bengal, Laccadive Sea, Andaman & Nicobar Seas and Indian Ocean region. We have exchanged the data / information under IOC/IODE network of NODCs, RNODCs and WDCs.

3. IODC Marine Data & Information Management System for the Indian Ocean

A user friendly system for the management of marine data /information has been developed for processing and retrieval of physical, chemical, biological, geological and geophysical data sets with common inventory database. It provides geographical, seasonal and parameter wise catalogue information and retrieval of data using multi parameter selection criteria. The system incorporates database for temperature, salinity, oxygen, nutrient, primary production, chlorophyll and zooplankton biomass in the water column and zoo-benthic biomass, micro-organism, geo-chemical parameters, marine magnetics, gravity etc at/in the bottom of the ocean.

4. IODC Projects on data management (currently under taken) are:

- 4.1 National Marine Data centre for oceanographic parameters sponsored by Dept of Ocean Development, New Delhi, India
- 4.2 Ocean data management for the Bay of Bengal Process study
- 4.3 To develop Marine Integrated Information System on the Indian Ocean (MIIS)
A collaborative work between IODC and All-Russian Research Institute of Hydrometeorological Information-World Data Center (RIHMI-WDC-B), Obninsk, Russia
- 4.4 A collaborative project on data management to assist NODC /NARA Sri Lanka.

5. Description RNODC-INDO and co-ordination:

RNODC-INDO is working to improve the oceanographic data and information services in the region. The center is also helping IOC member states by providing training to the data personnel on oceanographic data / information management in the IOC/INDIO region. The center is working to enhance IOC regional capabilities to interpret and to use results from field experiments through participation in the IOC regional programs. The center holds the oceanographic data sets for more than 70,000 stations collected from the Indian Ocean since 1900.

6. Data Analysis:

The following data sets were analyzed by the IODC staffs and published the results in research journal or presented in the workshop etc during 2001-2003:

- Data on primary productivity was processed within EEZ of India and fishery potential was calculated and the result was reported in the workshop.
- Quality control system for biological oceanographic data in the Northern Indian ocean was developed and the result was presented in the COD conference held in Brussels, Belgium.
- XBT data from Antarctic water was processed to check the probes fall rate in the Antarctic water. The result show that probe's fall rate is related due to the low temperature water in this region. Manuscript on the find of this work is reported to the Journal of Atmospheric and Oceanic Technology.
- Digital Bathythermograph (DBT) data was processed to check the systematic bias in temperature and the result reported to Journal of Oceanography.

7. Users interaction(Workshops/Training/Meetings):

IODC staffs participated in two workshop,two training course and two meeting on different topics related to marine science during 2001-2003. These are:

- Participated in the national workshop on bay of Bengal monsoon experiment (BOBMEX) initial results held at NIO,Goa,15-16 February 2001
- Participated in the international conference on the colour of ocean data (COD) held at Flanders Marine Institute ,Brussels,25-27 November,2002.
- Participated in the IOC Mission to Sri Lanka, to assist NODC, Sri Lanka, Colombo,4-10 November,2001.
- Individual training on oceanographic data /information management was provided to staff member from NODC Sri Lanka at IODC during July-August 2002
- Participated in the Indian Ocean Argo Implementation meeting held at Hyderabad ,26-27 July,2001
- Participated in JGOFS data management Task Team Meeting held in USA 28-31 January, 2002

8. Data products and service developed:

The following oceanographic data / information products /software /service were developed by IODC:

8.1 NIO-BIO-CD:

NIO-BIO-CD was developed by IODC, contains 3956 historical biological profiles collected in 1951 to 1996 on primary productivity, Chlorophyll-a, zooplankton biomass, zoo-benthos biomass & density and micro biological bacterial analysis in the Indian Ocean. A part of the data of 1194 profiles were collected by the institute on board the research vessels INS Darshak, R.V. Gaveshani and ORV Sagar Kanya during 1974 to 1996 and 2762 profiles were taken from the IOC data report (Published in 1974) having the biological data sets collected by many countries in Indian Ocean since 1951.

8.2 Data Visualization S/W:

A window based system has been developed for the selection , retrieval, and visualization of biological data archived in the IODC data bank. The system is developed in Visual Basic (VB) and it has three modules namely selection, visualization and retrieval of biological data sets. The first module for data selection operates on input information namely (i)area (range of lat. & long.or clicking the four coordinates from the map), (ii) Cruise / station references, (iii) Period (year / months) selection. The selected data profiles can be plotted by visualization module. The module also plot average profile of the variable considering all the stations of the cruise and overlay a profile of the station. Thus the module can check the variation in the station profile over the average cruise profile. The system has been developed, interfaced and tested with forty five years historical biological data sets.

8.3 BIO data quality control S/W:

A quality control system has been developed and performed the quality control(qc) check on the primary production(pp), and chlorophyll-a (chla) profiles in the Northern Indian Ocean collected during 1976 to 1997 by R.V. Gaveshani cruises(1976-94), and ORV Sagar Kanya cruises (1985-1997).These quality checks are based on the Meta data inventory, duplicate, range, maximum value of specific production, tail end increase and statistical. .Preliminary quality check is performed by visual inspection of the vertical profiles of pp and chla along with temperature. individual station profiles of the parameters were selected and plotted in the same plane so as to compare the nature of the profile easily. Other quality check on pp was done by computing the chla specific pp. Statistical check were carried out by grouped the data in $1^{\circ} \times 1^{\circ}$ Or $2^{\circ} \times 2^{\circ}$ Latitude-longitude square depending on the number of data sets available .Mean and standard deviation (SD) for the values from each depth were computed irrespective of the season. Maximum and minimum value for each depth were fixed as mean $\pm 5SD$ when land area contained in the square. In other case area the maximum and minimum value were fixed as mean $\pm 3SD$.Below 50m depth the criteria set was that the value should be between mean $\pm 3SD$.

8.4 On line hydrographic data service:

The on line hydro graphic data service has been developed with user friendly data selection and retrieval interface software. The hydrographic data profile on temperature, salinity, oxygen, nitrate, nitrite , phosphate, ammonia, silicate, Ph and alkalinity collected during 1906 to 1996 are available on the Indian-ocean web server of the institute for on line dissemination
URL <http://www.indian-ocean.org/support/main.htm>

8.5 Data announcement:

Two directory interchange format (DIF) were written for the Cds developed by IODC namely JGOFS(India)-CD and NIO-HYDRO_CD. And posted on NASA web site in GCMD for the users interaction and promotion.

Data announcement (DA-17) for NIO-HYDRO-CD was developed and distributed to the users under the user's interaction program.

9. Data Dissemination:

The following data /information were provided on request:

- A total 120 data requests were handles during 2001-2003 and the requested data were disseminated to users on computer media such as floppy, cds, email etc.
- Progress report on JGOFS data management were developed and presented to International JGOFS
- Oceanographic data for physical, chemical and biological parameters were processed for Arabian sea and bay of Bengal for the sponsored project.
- Provided training to a staff member from NODC Sri Lanka
- Twenty NIO_HYDRO-CD were distributed on request to the users.

10. Future plans:

IODC will further developing/ strengthening the oceanographic data/ information services:

- To develop WEB base data/information dissemination system
- To develop meta data directory for the Indian ocean
- To develop coastal data / information system
- To strengthen data /information monitoring system
- To develop value added information products for Indian Ocean
- To develop oceanographic knowledge base

11. To improve Ocean data/information management activities in developing countries, require:

- To develop Infra structure facilities such as high speed Communication link, H/w, S/w, Human resource Development (HRD) etc

- To develop capacity building required Special efforts
- To study the coastal processes required information/data in the coastal region .
- To develop Projects on data analysis and interpretation
- To develop value added data products.

12. IODC Address and URL:

Head, Data & Information Division
National Institute of Oceanography
Dona Paula Goa ,Pin 403004 India
Tel. 91-(0)832-2450211
Fax. 91-(0)832-2456702/03
Email saruj@nio.org
Url <http://www.indian-ocean.org/support/inodc/index.html>

Project data management in BODC -- Roy Lowry

1) BODC Modes of Operation

In addition to its role as the UK National Oceanographic Data Centre, BODC has been offering project and operational data management services since 1988. The activities undertaken in each of these modes is summarised in the following sections.

1.1) NODC Activities

The primary activity within the NODC is the ingestion of data into the UK National Oceanographic Database (NODB). Data, generally moored instrument records and CTD data, are routinely received in batches from organisations such as the Fisheries Laboratories (CEFAS and FRS) and the Southampton Oceanography Centre. The data received are converted into BODC's working format, documented, quality controlled and loaded onto the base.

Further work is being undertaken to document and quality control the historical data from the UK National Tide Gauge Network for incorporation into the NODB and inclusion into an on-line delivery system that currently supplies data from 1990 until 3 months ago.

Users are served by the NODC through web-based metadata and data systems plus a Requests Officer service. The human element allows referrals to other sources of information and manual integration of the holdings of the NODB and project databases to satisfy users' data requirements. An important part of BODC's NODC activities are interactions with other national and international organisations concerned with marine data management such as the BODC-co-ordinated UK Inter-Agency Committee for Marine Science and Technology (IACMST), ICES, IOC/IODE, GOOS/JCOMM and the World Data Centre Network. These interactions include routine data submissions to both the WDCs and ICES.

In its NODC role, BODC participates in a number of European Union data management projects. The current project portfolio includes the pan-European metadata networking projects SEASEARCH and EDIOS, and the E-Seas sea level network.

1.2) Project Data Management Activities

BODC also provides data management services to active oceanographic research projects through a form of working termed Project Data Management. This includes input into project planning, active project participation and project data set publication, usually on CD-ROM. A detailed description of BODC's project data management activities is given below.

1.3) Operational data centre

Near real-time or operational data management services have been provided to users of the national tide gauge (NTG) network since **1989** and have been developed over the past 2 years for the UK Argo floats.

The NTG data are retrieved weekly and are made immediately available in their raw state to operational data customers. The quality controlled data are freely available as part of a delayed-mode data set through a web-based system 3 months after collection.

BODC routinely receives daily satellite messages from all UK Argo float deployments, which are unscrambled, assembled into profiles and made available on the web as quickly as possible. Quality control work then commences to assemble the final delayed mode data set. Routine submissions of data in Argo-specified formats are made to the international Argo data centres.

In the case of Argo, the boundary between project and operational data management becomes blurred or even totally artificial. There are strong indications that the handling of real-time data will become increasingly important in project data management.

2) BODC Staffing Compliment

- Directorate
 - Director and group administrator
- IT infrastructure and development
 - TD, DBA, webmaster, 4 developers
 - 2 short-term parameter dictionary developers and a placement student
- NODC
 - Deputy Director and 4.5 data scientists (1 vacancy)
- Project Data Management
 - 7 data scientists (2 vacancies)
- Operational data management
 - 2 data scientists
 - Part-time systems developer

3) Project Data Management

3.1) Definition and Scope

Project Data Management is the provision of data management services to active research projects and is sometimes known as 'End-to-End' data management. There are a number of facets to the services provided:

- Input to project and cruise planning, which ensures data management issues are addressed (or at least not totally ignored) at the planning stage.
- Cruise participation as data managers, which ensures that all metadata accompany the data off the ship, gives data scientists a better understanding of the data and forges working relationships between scientists and data managers.
- Data processing. Some data are taken straight off the ship, worked up by BODC and distributed to project participants. This has several advantages:
 - Multidisciplinary data handling is co-ordinated.
 - Infrastructure replication is avoided.
 - A guaranteed minimum data quality standard is provided. This may be exceeded in some cases where specialists undertake part of the work for a project, but it is always maintained.
 - Data exchange "currency" is provided to encourage submission of sample data sets to BODC.
- Project participation. BODC data scientists attend and present at project science workshops, which keeps data managers aware of the science and provides an opportunity to apply "name and shame" data delivery pressure.
- On-line data delivery. A revolutionary (in 1989) line-mode and native SQL interface was provided and used for several years. However, no resources were available for its development and it fell into disuse. A replacement web-based system is now in place and due to be enhanced by software development work currently in progress.
- Project data set publication on CD-ROM, providing a value-added data set through BODC's QC and documentation procedures and a concrete deliverable for the project.

3.2) BODC Project Data Management History

The concept of Project Data Management was developed in BODC in response to our being asked to manage the data from the NERC North Sea Project. This ran from 1988 until 1990 starting with 28 back-to-back cruises in 14 months. BODC were asked to take on data management 3 months before the first cruise.

Project data management “invented” and the infrastructure (100,000 lines of FORTRAN, some of which is still running) was rolled out in 6 months. Initially, this was viewed as a ‘one-off’ exercise, but other projects were soon to follow.

The first was the Biogeochemical Ocean Flux Study (BOFS), which ran from 1989 until 1992. This was based in oceanic waters rather than a shallow sea, which brought new data-processing problems and a raft of new parameters. However, the systems developed for the North Sea Project were amended and adapted and the resulting set of patches just about coped.

In 1993, BODC was asked to provide Project Data Management services to the European Union Ocean Margin Exchange (OMEX). At this stage it was realised that Project Data Management was becoming a permanent feature of the BODC portfolio and that OMEX, involving scientists and research vessels from 10 nations, was on a different scale to anything managed to date. Consequently, the systems were significantly re-engineered, forming the basis of those currently in use.

Since OMEX the principles of Project Data Management have been applied by BODC to the management of data from the following projects:

UK WOCE	Arabesque	LOIS (RACS and SES)
PRIME	ACSOE	PROVSS
DISCO	PROPHEZE	Autosub Science Missions
INDIA		

3.3) BODC’s Current Project Portfolio

BODC is currently working on data from a number of NERC projects including:

- Marine Productivity
- Marine & Freshwater Microbial Biodiversity (M&FMB)
- Atlantic Meridional Transect (AMT)
- POL Coastal Observatory
- Microbially-Driven Biogeochemical Processes, Exchanges and Controls (MDB)

In addition, negotiations are underway for BODC to take on the data management for the Rapid Climate Change and UK SOLAS projects.

3.4) Relationship between Project Data Management and the NODC

When the North Sea Project was viewed as a “one-off” activity the original plan was that once the project was completed the project data set would be assimilated into the NODB. However, this did not happen. Consequently, the situation has arisen where BODC has two large-scale data resources that may be described as “loosely coupled,” but certainly not as integrated. It has now reached the stage where the data holdings of both systems are of a similar magnitude. Only the CTD data are exposed to any sort of data discovery system.

A number of parallel initiatives are currently underway to address these problems. Resources are now available in BODC’s NODC operations to migrate some of the project data onto the NODB, with priority directed at the 300+ cruises of worked up underway data. Secondly, a web-based interface is under development to allow free access to ascertain what the databases hold and authorised access to extract the data. In the longer term it is planned to use the technology being developed in the NERC DataGrid project to provide seamless interoperability between the project databases and the NODB.

3.5) Possible Future Developments

There are two areas that are either outside or just within BODC's current portfolio that will become increasingly important for our future project data management activities. It is clear that there will be an upsurge in operational services as more and more projects are utilising moorings or tethered profilers that telemeter data in real time. It is also clear that BODC will need to develop additional services to manage model data, either for "definitive run" stewardship or as part of operational systems.