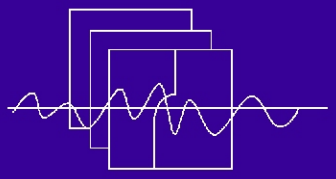
Vol 15 No 3 - October 2004

SADCO SADSO



Southern African Data Centre for Oceanography
P O Box 320, Stellenbosch 7599
South Africa

Email: mgrundli@csir.co.za

Website: http://sadco.csir.co.za/

SADCO is sponsored by ...

Department of Environmental Affairs & Tourism
SA Navy
CSIR Environmentek
NRF (SA Universities)
Namibian Ministry for Fisheries & Marine Resources

Real time data: changing roles for data centres?

In the past, data centres could be compared to somewhat dusty places, continuously being loaded with data in all shapes and sizes (formats). It was the last thing on the checklist of oceanographic projects (submission of the data to a data centre), and as such at the "end of the road" of the flow of data. On occasion, it would pass some of its contents to interested users. The data was mainly historic and because of its age largely drained of useful information. As such, the data centre's role was to archive and safeguard the data for future users, much like one of the functions of a museum.

At the time, oceanographic programmes took years to plan, and even more years to collect and process the data. The "time scale" of the data centre was therefore somewhat in line with the flow rate of the various data streams.

Over the years, this picture of a data centre has been changing, and the rate-of-change is continuously accelerating. A main driving force in this transformation is the increase in the very rate-of-change of the environment itself, and the need to keep track of, manage, and even predict these changes. The corresponding aim of many oceanographic programmes is to produce data closer to real time. Satellite programmes are a prime example, where the developments in the data collection technology not only expedited the availability of such data but have cascaded into downstream analyses methodology and into models that are now able to assimilate the data on a global scale in near-real time. This is another (major) step toward oceanographic *forecasting*.

The appropriate management of data, and specifically a smoother integration of data management with data analyses and interpretation is obviously called for. The ultimate success of the analysts and advisors is keenly dependent on having the right data at the right time in the right place.

This Newsletter reminds readers of some regional/local developments in the framework of the "real time" oceanographic field, specifically where data management plays a strong role. It also tries to indicate the changing role of a small data centre such as SADCO on the African continent, where some strategic re-alignment of its outlook may be required. In the next Newsletter, this "introduction" will be continued, in the light of some relevant Workshops that have been held in October and November 2004.

ISSN 1017-2351

OBIS Regional Nodes meeting

SADCO's foreseen involvement in OBIS (Ocean BioGeographic Information System) has been indicated in a previous Newsletter. The OBIS structure will consist of an "OBIS Central" at Rutgers University (see Fig. 2), as well as a number of regional OBIS nodes, or RONs, distributed over the world. One of these, the Sub-Saharan Node, will be operated by SADCO.

The first RON meeting was held at the Bedford Institute for Oceanography (BIO) in Dartmouth, Canada in October. Because SADCO's manager was out of the country at the time, **Ursula von St Ange** was invited to attend the meeting.

The main objective of the meeting was to initiate the 2 year project to develop a global network of Regional OBIS Nodes (RONs). There were 23 meeting participants (see Figure 1) from 14 countries, representing 8 of the 10 currently recognized RONs.

The meeting consisted of oral presentations and plenary and working group discussions about

- OBIS's current situation
- it's future
- technical issues about the main OBIS site and
- hosting an OBIS database and portal at the regional nodes.

The following are some brief issues emanating from the meeting.

Loading data into OBIS

- Regional nodes will capture (or facilitate capturing) relevant information.
- OBIS Central (Rutgers) will periodically upload data from the RONs, to populate a Common OBIS Data Set (see Fig. 2), making use of database crawlers (written with Java and JDBC).

Figure 1. Attendants at the RON meeting, Canada.

Front Row:
J. Black, D. Ricard, M. Lewis, K. Stocks, U. Von St. Ange, V. Chavan, F. da Silveira.

Back Row:
P. Zhang, P. Halpin, A.Marino, R. Branton, M. Costello, T. Rees, F. Grassle, R. Froese, R. Escribano, M. Kennedy, D. Robertson, L. Van Guelpen.

Missing:
P. Boivin, S. Song, G. Pohle, K. Zwanenburg

OBIS Regional Nodes meeting (continued)

Extracting data from OBIS

- Originally OBIS requests were handled via a search of distributed data providers, which could take several minutes at a time. As the number of data providers and data increased, the search time became unacceptably long. In 2003-4 the search action was modified to incorporate a central "OBIS Index" permitting spatial searching and production of "quick maps" and a central "OBIS Cache" to improve retrieval speed.
- OBIS searches are now structured into a 2-stage process:
  - "Stage 1" searches, which operate on the Index, can return information on up to 500 species very rapidly (e.g. <5 secs), a vast improvement on the original OBIS system (several minutes per search, with some or many searches returning no data).
  - "Stage 2" searches, which retrieve relevant point data for any species from the Cache.
- All nodes will access a common OBIS data set and 3rd party analytical tools at a central location. There are two mapping tools available to the users. The c-squares mapper from the CSIRO Marine Research in Hobart, Tasmania, and the KGS Mapper from the Kansas Geological Survey.

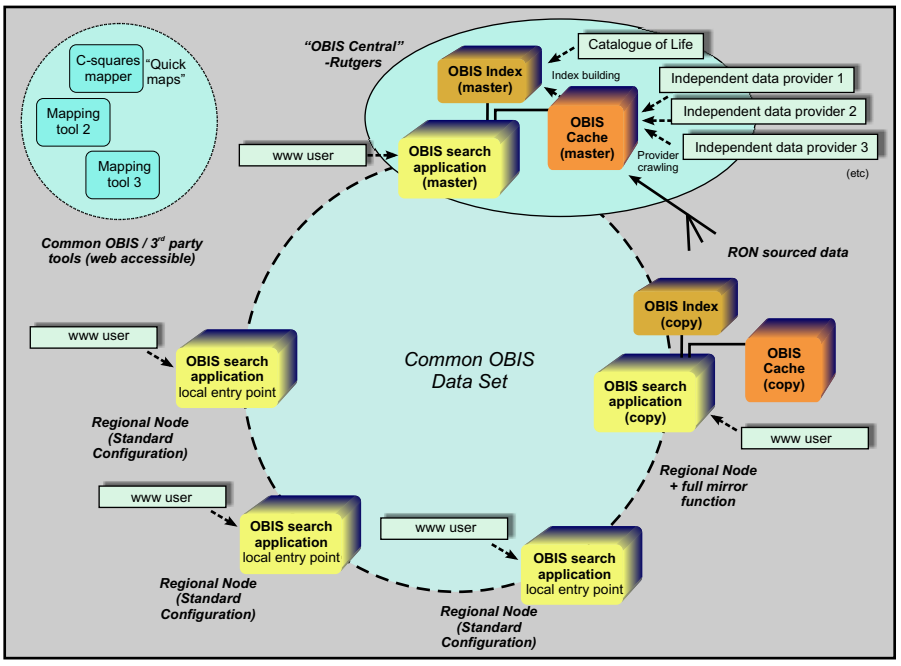


Figure 2. Schematic diagram of OBIS framework [from OBIS Management Committee Meeting Report (OBISMC1), September 2004.]

Starting the development of RONs

As all RON portals will have the same look-and-feel as the main OBIS node, there was general agreement to initiate development of regional start-up kit. This would start with development of simple prototype portal (using 'Plone' portal manager) modelled after the existing OBIS Portal. The RONs will be able to incorporate the ability to customise the search entry point to suit local needs.

The OBIS schema has additional data fields for marine data, so will be better suited for ocean biodiversity data than GBIF (a motivation to rather contribute towards OBIS; all OBIS data is uploaded into the GBIF database anyway).

Integrated Ocean Observing System

The Southern African oceanographic community is familiar with GOOS (Global Ocean Observing System), and probably of its regional "subsidiaries" spread all over the world, such as Euro-GOOS, MedGoos, GOOS-AFRICA, IOGOOS, etc (see http://ioc.unesco.org/goos).

In an article in the Spring 2004 edition of *Earth System Monitor* ("US Integrated Ocean Observing System" or US IOOS, S Hankin, T Malone and R Cohen provide insight into the way GOOS is handled in the United States.

Extracts from the IOOS description

The SADCO readers may benefit from brief extractions of some points from the article, to highlight some corresponding aspects of the southern African scene. Interested readers are referred to the article itself for more comprehensive information.

1. The instruction to design and implement US IOOS came from the US Congress.
2. IOOS will deal with past, present and future states (of the oceans).
3. It will address the following areas (from an oceanographic point)
    a. Climate variability
    b. Safe and efficient marine operations
    c. National security
    d. Sustainable use of resources
    e. Healthy marine ecosystems
    f. Mitigation of natural hazards
    g. Public health
4. IOOS will comprise three subsystems
    - Observations
    - Data management and communications. This system is considered by the authors as the "primary integrating element of IOOS".
    - Modelling and analysis

The Data Management and Communications subsystem (DMAC)

- For the DMAC to be effective, independent data management systems will need to be connected.
- In a novel move, data discovery (searches) will also make use of commercial web search engines (such as Google and Yahoo) (apart from own software).
- Designated Archive Centres will make data available on-line, as far as possible, at no cost. Some service cost will be levied for data from off-line sources.
- Plans will exist to ensure that data can successfully be moved to other platforms when the need arises.

Although it is felt by the authors that the technology required for IOOS can be developed, a significant challenge is recognised in the coordination and cooperation between IOOS partners and user communities.

So much for the summarised extractions from the EMS article.

Southern African relevance

- The areas addressed by IOOS (Item 3 above) are closely linked to the outcome of WSSD, Johannesburg, 2002 (see e.g. Implementation Plan Items 29, 30, 33, 36, 40, and 42). These hold equally for southern African programmes on a regional, national or local scale (incl. GOOS-AFRICA).
- GOOS AFRICA is presently chaired by Justin Ahanhanzo (j.ahanhanzo@unesco.org) and the incoming chair of the GOOS Steering Committee is John Field (University of Cape Town) (Prof Geoff Brundrit, pers. com).
- GOOS AFRICA has taken the regional view and linked itself to regional groups like the BCLME. The need is obviously for an arrangement that is not project-driven and has greater longevity.
- Both Justin Ahanhanzo and Tom Malone (the latter co-author of the EMS article referred to above) attended the BCLME (Benguela Current Large Marine Ecosystem) Workshop 8-11 November in Cape Town. One of the sessions of the workshop was devoted to "Maritime Operations in the Benguela coastal ocean", where various industries indicated their need for forecast metocean and other operational information.
- One of the papers in that session introduced a scalable modelling approach, focussed on mesoscale hubs (areas where maritime activities, shipping, etc have a required "density").

How is this related to SADCO?

a) SADCO seems to be the only "permanently networked", operational oceanographic entity of its kind in Southern Africa. Its funding and management model is indicative of its multi-organisational and transboundary nature.

b) It is most likely that the Southern African marine and metocean organisations that would participate in GOOS AFRICA are already represented to a certain extent on the SADCO Steering Committee. The fruitful role that it can play in this regard has been recognised by the Steering Committee itself. It is suggested that, where opportunity and funding allows, SADCO becomes more closely involved with GOOS AFRICA.

c) SADCO has a target area that extends well beyond the countries that actively participate in SADCO. This would allow it to contribute to large programmes, such as BCLME and the foreseen Agulhas- Somali LME.

d) SADCO is progressively moving into additional oceanographic domains, as opportunities and sustainable funding arise. In this way, its disciplinary "footprint" is continuously enlarging.

e) Organisations where large amounts of metocean data are handled in an on-line, real-time mode, are members of the SADCO consortium.

F) One of these organisations that plays an important role in the metocean data stream, the South African Weather Service, is also part of the SADCO Steering Committee.

In conclusion it therefore seems that if logistic and other aspects can be resolved, SADCO is well placed to play a meaningful role in some aspects of GOOS-AFRICA.

Oceanographic metadata and raw data

In a recent article "Introducing the US NODC Archive Management System, by DW Collins, SB Rutz, HL Dantzler, EJ Ogata, FJ Mitchell, J Shirley and T Thailambal (*Earth System Monitor*, Vol 14 (1), 2003), insight was provided in the structure and functioning of the Archive Management System at NODC.

The article indicates the various components that make up a "survey" (or data submission), namely, metadata, accession number, descriptions and keywords, virus checks, actual data, etc., and how they are linked.

A brief description of how SADCO handles aspects of metadata vs raw data would be appropriate, to indicate some of the aspects of local (oceanographic) data management.

SADCO Inventory

The SADCO inventory (sadco.csir.co.za), which is accessible to all users, provides information on all surveys to which SADCO has been alerted. In this way, SADCO's inventory resembles an on-line, searchable version of the CSR (Cruise Summary Reports) system.

The only criterion for inclusion in SADCO's inventory is that the survey must have been located within SADCO's geographic target area (10° N to 70° S, 30° W to 70° E).

The way that entries for the Inventory are obtained are mainly:

- Information supplied by the collector (chief scientist, or organisation)
- Scouting of published material for relevant surveys (e.g. list of cruise reports from a foreign oceanographic institute, or journal article)
- Notification of an upcoming survey to be conducted in SADCO's target areas, via a state department or other authority.
- Metadata record created from an actual submission of raw data.

When an entry is made on the Inventory from information for which actual data has not been submitted, the Inventory is checked for a duplicate.

When actual raw data is submitted to SADCO, the data is linked to an already existing entry in the Inventory, or a new entry is generated.

The result is that:

- For every data set in the Data base, there is a corresponding entry in the Inventory
- Every entry on the Inventory does not contain a corresponding data set in the data base.

The Inventory therefore contains three types of entries:

- Type A: data has been submitted to SADCO
- Type B: data has not yet been submitted to SADCO
- Type C: data will not be submitted to SADCO because the data type lies outside SADCO's disciplinary domain.

So what is the unique link between information in the data base and the Inventory?

This link is the **SurveyID**, *a unique alphanumeric identifier* associated with each survey. Using this parameter, cruises can be located uniquely in the Inventory as well as in the database (see below). Other descriptors, such as vessel, chief scientist, or date, are not unique, and therefore do not allow location of a single survey.

Flagging of data

Data flagging has been introduced for two reasons

- Protection of **confidential information**. This arises from researchers requesting time to complete analysis, publication, theses, etc, before data is released. SADCO provides the possibility for data to be flagged for 2 years after submission (i.e. not after collection), and this can be extended another year on request.
- **Quality aspects**. Upon request by the data donor, or through internal QC routines run by SADCO, data can be flagged because the quality may be questionable.

All flags are attached at station level (not survey level), allowing individual stations to be cleared.

Flagged data will only be accessible to a user who has access to the keyword provided to the initial data donor.

Scouting for "missing" surveys

SADCO uses Type B entries in the Inventory to scout for data. The owners of the data are then contacted to request the data itself.

Additional data for an existing survey

Data submitted for a survey may initially consist of bottle data only. Later, other data sets may follow (CTD, XBT, ADCP, etc). These additional data sets are loaded onto the database under the same SurveyID. So, knowing the SurveyID will allow extraction of all data related to a particular cruise.

Each data set (e.g. CTD, XBT, etc.) is identified with a Descriptor, to indicate the data type. In this way, the user will know if a particular temperature was collected with an XBT, with a CTD, etc.

Extraction of a specific data set (e.g. XBT) is not possible from SADCO's www front end, but off-line extractions can be customised around a particular data type, if requested.

Extraction of data for a specific survey

The method normally followed by a user to identify and obtain a specific cruise, is as follows:

Using the Inventory, the cruise is located by alternatively searching according to

- The *vessel*. If the vessel had operated for a longer period, the years would be indicated and the user can select a specific year
- *Date Chief scientist*. This information may not always be known
- *SurveyID*. This is the unique identifier (and normally the item being sought). However, if the SurveyID is suspected, searching on this item will confirm that the right cruise has been selected.

Once the right survey has been identified (and the corresponding SurveyID determined), the cruise data can be extracted from the data base.

Road forward

The ESM article referred to above mentions the significant effort required to move information from one system to another. SADCO can vouch for that (albeit on a much smaller scale), as evidenced from the transfer of SADCO from Informix to ORACLE.

One of the issues addressed by the US NODC archiving system is the role that the new system will play in implementing an integrated ocean observing system, where an increase in data flow is expected. This is also mentioned in another article in this Newsletter on the US-IOOS.

New US web site with near-real time global data

The ECCO web site (Estimating the Circulation and Climate of the Ocean, http://www.ecco-group.org) uses a general circulation model to assimilate satellite and other data, and provide global products at 10-day intervals.

SADCO III started at a time (1990) when the world was waiting for the launch of TOPEX/POSEIDON, and before the promise of high quality, routinely delivered, global data became a reality. While initial satellite data sets were accessible only in a hind-cast mode, the producers of satellite information have over the years moved ever closer to real time.

Where are we today? Looking back over the past 10 years the progress has been phenomenal: Data is being collected, by a number of satellites, received (by various ground stations), processed and corrected (by a number of organisations), assimilated (together with other data) into a global circulation model, and distributed world-wide in a time span that is now becoming real time (officially still designated as "near real time").

While the data streams from satellites, especially the altimetric ones, are in "near real time" mode already, it must be remembered that these streams will have a non-global coverage if "near real time" means data not older than, say, 5 days). To improve the geographical "coverage" other assimilation methods must be used.

For the climate researchers and global oceanographers among the SADCO readers an insight into near-real time information can now also be obtained from a web site of the consortium for "Estimating the Circulation and Climate of the Ocean" (ECCO). Analyses are done at 10 day intervals and are available for more than 10 years. The general circulation model developed at MIT is used (horizontal resolution of between 1° and 1/3°, and 46 vertical levels). The analysis assimilates satellite altimetric data (TOPEX/POSEIDON and Jason-1) and temperature data from the Global Telecommunication System (http://www.ecco-group.org and http://ecco.jpl.nasa.gov/external). The user can select from a variety of parameters, dates, and depths, and obtain colour plots within a few seconds.