Vol 15 No 2 - July 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

OBIS: Upcoming event

OBIS (Ocean Biogeographic Information System), the marine component of GBIF (Global Biodiversity Information Facility), is starting to gain momentum.

Brief background

SADCO provided inputs to a proposal for international funding, to host the Sub-Saharan node for OBIS (for more information and background, see SADCO Newsletter March 2004). The 2-year funding should allow the facility to be created, and available data to be loaded. The initial feedback for the funding looks very promising.

OBIS meeting in September

A meeting has been called for all the Node managers, and this will be held at the Bedford Institute for Oceanography in Nova Scotia on 18-20 September 2004.

Because of personal commitments, the foreseen node manager (Marten Grundlingh) will not be able to travel to Canada at that time. However, it is expected that there will be some important discussions on technical aspects of the nodal architecture, accessibility, formats, the network, etc. It has therefore been decided that there will be distinct benefit if **Ursula v St Ange** attends the meeting. Ursula has been a key role player in the design and construction of the various data bases that house SADCO's data. Discussions and decisions on these topics will therefore be right in her domain.

Ursula v St Ange

It is hoped that there will be the opportunity during, or before/after the meeting to interact with specialists at the Bedford Institute, to get some insight into the practical side of the node design.

ISSN 1017-2351

Underway ADCP data: update

As indicated in the Newsletter of November 2002, Marine and Coastal Management has more than 140 cruises on which ADCP (Acoustic Doppler Current Meter) data was collected in an “underway” mode. These cruises were done on the *Africana* and the *Algoa* since the 1990's.

On many (most?) cases, the velocity profile was collected while the vessel was on station (e.g. for CTD casts). Considering the shear number of CTD stations that are done on the cruises where ADCP data is collected, it is estimated that the total number of ADCP profiles that are available, is quite large.

In an environment with strong currents and changes in flow patterns, but a general scarcity of data, the value of the ADCP data could be quite significant. The cruises extending to the area of the Agulhas Bank and capturing the inshore edge of the Agulhas Current could provide insight into the many interesting and dynamic features of the Current.

The processing and loading of this data has been an item on SADCO's work list for more than a year. Only once the data has been processed and displayed, can the true value be estimated (also considering the quality of the data itself).

To unlock this potentially valuable data set, suitable software and manpower is required.

A processing package has been purchased by MCM, but there is presently no staff available at MCM to undertake the processing.

In the mean time, in anticipation of the forthcoming data, the loading and the extraction software has been written. This will allow a speedy loading of the data once the data stream has been established. Suitable products will be created after some data has been loaded, depending on the foreseen uses.

F.R.S. Algoa

F.R.S. Africana

Loading of moored ADCP data

As has been reported in the December 2003 newsletter, moored ADCP data has been obtained from Mike Roberts (Marine and Coastal Management). The data has been collected as part of the Coelecanth programme.

The data has now been reformatted and loaded.

The data will remain flagged (= restricted access) for a period before it can be made available to users.

[*Just a note on SADCO's flagging policy: SADCO's constitution allows donated data to be flagged for a period of up to 2 years after submission, extendable by another year. Access to such data is restricted (password protected) to the data donor only. However, if users want to obtain insight into means over a given area, such a request can be submitted offline to the data centre, and flagged and unflagged data will be used to construct the means.*

The flagging of data is considered a compromise with data donors, to ensure that data is submitted to the data centre expeditiously while at the same time not jeopardising the donor's ability to publish the data. By timing the flagging period from the date of submission, rather than collection, it is believed that ample opportunity is given for donors to reap appropriate benefit from the data.]

Cleaning up the marine data base

It is SADCO's policy that data submitted to the data centre for safekeeping and dissemination should be fully edited and quality controlled (QC) before submission. The reasons for this are

a) SADCO does not have the local insight into aspects of data quality that would reside with the data collector. E.g. part of the quality control of data is the calibration of equipment, checking for anomalies, data drift, spikes, and adding suitable "header" data (position, time), etc., and this is the responsibility of the data collector.

b) To undertake proper QC takes time, diligence and commitment. This translates into funding, and SADCO does not have the resources and financing to undertake this task.

It seems, however, that not all the data submitted to SADCO in the past has been fully "cleaned". In spite of the fact that a degree of checking is done during the loading process, it is virtually impossible to do all checking for all the parameters. By having errors in the data, it reflects negatively on the status of the data centre, and its ability to provide quality data.

In addition, when applying QC procedures, finding an error is one thing, replacing it with the "correct" value takes a hundred times longer.

What sort of errors are we talking about, and how do they slip through?

Errors can occur in all parameters, and are sometimes not immediately apparent.

Those that handle data know that the list of things that can go wrong, is endless. In the age we live in, many of these "errors" are introduced during formatting and reformatting as data is passed from organisation to organisation.

Researchers typically handle less than 1000 stations per year, and it is possible to undertake detailed and extensive QC processes. However, this is not the scenario often faced by data centres. A few years ago, SADCO loaded about 50 000 stations in one year (obtained from international sources), and found it impossible to check individual values or even individual stations.

Can a data centre reject data of doubtful quality?

The short and theoretical answer is yes.

The longer and more practical answer looks different:

a) A data manager often finds him/herself in somewhat of a predicament: On the one hand, the data centre is keen to ensure that quality data is properly archived. On the other hand, the manager realises that the data collector is also facing financial and resource constraints, and does not have unlimited time in which to complete the data processing.

b) Most data centres operate within a national or regional framework, and have corresponding obligations in terms of data scouting and custodianship. The data centre is therefore obliged to prevent data becoming lost.

c) Often, researchers extract only a portion of the data from the collected set, and leaves the rest for "later checking and analysis". However, the time scale of this "later checking" often starts approaching the period of practical data recovery (hardware and software platforms change with time, and data stored on older media becomes difficult to recover). The data collector is suddenly forced to transfer the data to a data centre before having had time/opportunity to revisit the data. This is especially the case where the data collection is of a routine nature (repeat cruises, repeat lines, etc).

d) The initial data collector may have left the organisation, leaving (often unwilling) successors to do checking and processing).

e) Within the oceanographic community there is also an attitude that "bad data is better than no data".

So, in the end, a data centre tends to accept data of varying quality.

Because of its experience with data of uncertain quality, SADCO is upgrading its QC procedures. In the process, use will be made of protocols developed internationally. A brief insight into examples of errors that occur, and how they are identified, is given in the table.

Some errors that can occur in submitted data, and some identification methods

Error source	Example	How identified
Station position	Reversal of lat & long, duplicate lat/long	Graphic plot and vessel speed check.
	Station position is located overland	Check against topography
Coding and formatting	Atmospheric pressure is coded as 999 which is a valid reading, in stead of 9999 which is a null reading (= no measurement).	Programmatically
Depth	The deepest measurement is taken at a depth deeper that the expected bottom depth	Check against topography. Possible incorrect position
Profiles (temperature, salinity and oxygen)	Values too high or too low	Use World Ocean Atlas2001 and 5 degree annual average for climatology envelope
	The T, S and O_2 profiles contain spikes	Identified via a spike check.
	Spikes in T/S profiles	IOC Manual algorithm
	T/S profiles incorrect gradients	IOC Manual algorithm
	T/S density inversion test	IOC Manual algorithm
	Reversals in the vertical profiles (Instrument moving inconsistently).	Lowering speed check.
Instrument calibration	Long-term change in data record (e.g. with deployed underwater instruments, such as current meters).	Graphically identified.
Header	Some older cruises had incorrect surface information (e.g. Atmospheric Pressure, Swell Height and Swell direction).	Check against WMO (VOS) limits
	Many of the older cruises do not have an instrument/data descriptor or are incorrectly identified.	
Time	Station times are occasionally mixed up between SAST and GMT.	Vessel speed check, intercomparison with other stations.

Cleaning up the VOS data

After designing a suitable set of quality control procedures for the VOS data (see Newsletter of September 2002), the application of these criteria to data already loaded in the data base, has been completed.

Firstly, it should be noted that the VOS database is huge. Within SADCO's target area there is a total of about 5 million records. Each record consists of at least one parameter such as air temperature, sea temperature, wind speed, wind direction, wave height etc. Most of the observations have a full suite of parameters, so that the estimated number of actual observations is approximately 50 million.

Secondly, it will be recalled that the errors that were identified were located in the outliers (maximum, minimum), so *it was expected to find errors in far less than 5% of the data.*

Mario August and Ursula v St Ange undertook the clean-up process, and the following summarises the outcome.

For clarification, the VOS data holdings are separated into the following data bases (indicated are the actual limits of the observations):

	Latitude	Longitude	Period
Vos_main	8°N - 80.8°S	0° - 49.9°E	After 1960
Vos-main2	7.2°N - 69.9°S	0° - 49.9°E	Before 1960
Vos_arch	10°N - 78.9°S	31°W - 69.9°E	After 1960
Vos_arch2	9.9°N - 69.9°S	35°W - 75°E	Before 1960

Mario August

The process comprised the following:

- The parameter limits were embedded in a programme that could be used to scan systematically through the VOS data base.
- Values falling outside these limits were removed to a separate "rejected folder"
- A total of **41091** values were removed from the data bases.
- The bulk (>90%) of the discarded values are from the dew point temperature and atmospheric pressure.
- The process was repeated to ensure correctness

The table below gives an indication on how many values are currently stored in the database.

Number of observations remaining

	Vos_Main	VOS Main 2	VOSArch	VOSArch 2	Total
Drybulb	2157010	865775	2321523	1298734	6643042
Dew-point Temp.	1835853	146609	1944300	135808	4062570
Sea-level pressure.	2131980	644862	2274042	935391	5986275
Swell height	1402555	230248	1385478	163932	3182213
Swell period	1276686	108570	1240783	229219	2855258
Sea Surface Temp.	1989461	781465	2171893	1226518	6169337
Wave Period	1528836	89169	1566680	417371	3602056
Wave Height	1723632	381448	1745634	354822	4205536
Wind Speed	2169897	905188	2278427	1347706	6701218
Wetbulb	1535098	208852	869867	153352	2767169
TOTAL	17751008	4362186	17798627	6262853	46174674
	38%	9%	39%	14%	100%

This means that about 0.1% of the data was removed.

The 46 million values are contained in the following number of records currently stored in the database.

Number of records

Vos_main	**2202388**
Vos_main2	**909549**
Vos_arch	**235031**
Vos_arch2	**1375811**
Total records	**4722779**

So what happens next?

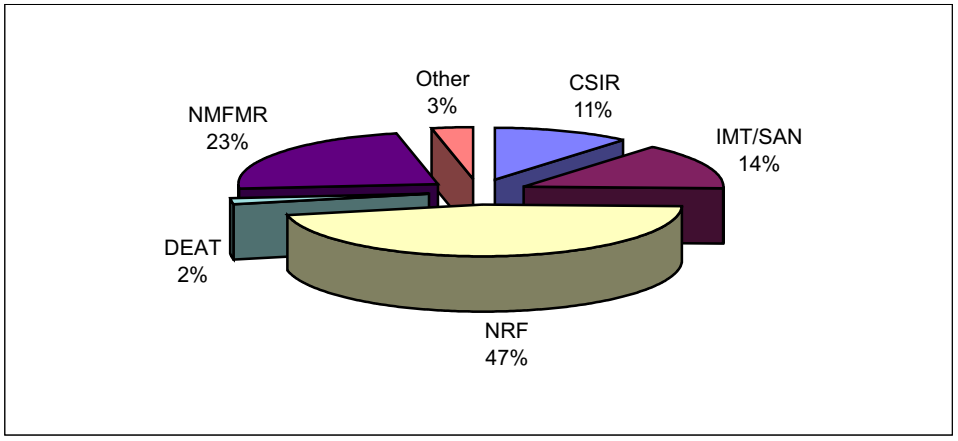
The new screening criteria for each parameter will now be built into the VOS loading programme, to ensure that values are filtered as they are loaded.

Offline requests

During the 2003/4 financial year, about 17% of SADCO's budget was devoted to handling off-line requests. This is down from the previous year's 28% (which was the highest off-line request load since 1990), but is still considered very high for a data centre.

The enclosed chart indicates the distribution of the request origin. The NRF and NMFMR (Namibian Ministry for Fisheries and Marine Resources) were the largest request users. The "NRF" usage is on behalf of students wanting information for research purposes (the main users were from UCT and UPE). "Other" requests in 2002/3 refer to those originating from the public.

Louise Watt handles most of the offline requests. Users are invited to make use of the on-line or off-line facility.



Distribution of requests submitted to SADCO

It should be remembered that the direct extraction of data is not the only yardstick for the general use of a data centre. The services that are directly to the benefit of users include:

- Archiving data (placing the data in a safe environment with long-term recoverability, taking over the responsibility for providing the data to a third party)
- Custodianship of data (includes the responsible and
- Systematic management of the data, flagging data to
- restrict access, regular back-ups, upgrading the soft and hardware, etc)
- Providing suitable, modern, on-line access systems
- Scouting for data from third parties
- Guidance on quality control
- Regular communication with users on data availability.

Louise Watt

Layout & Design: Magdel van der Merwe
CSIR