Vol 18 No 4 - December 2007

SADCO SADSO

SADCO Chairman receives PhD

Exactly a year ago the Newsletter announced the election of Raymond Roman as Chairperson of the SADCO Steering Committee. It was a great pleasure to the SADCO Steering Committee to be informed at its November meeting that Raymond would be graduating with his PhD at the University of Cape Town in December. We have since learnt that the degree was awarded at the graduation ceremony on 10 December.

Congratulations, Dr Roman!

***Raymond Roman**, in high spirits at the SADCO Steering Committee meeting in November 2007, after the Committee learnt about his imminent graduation.*

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

Manager: **Marten Gründlingh**

Email: **mgrundli@csir.co.za**

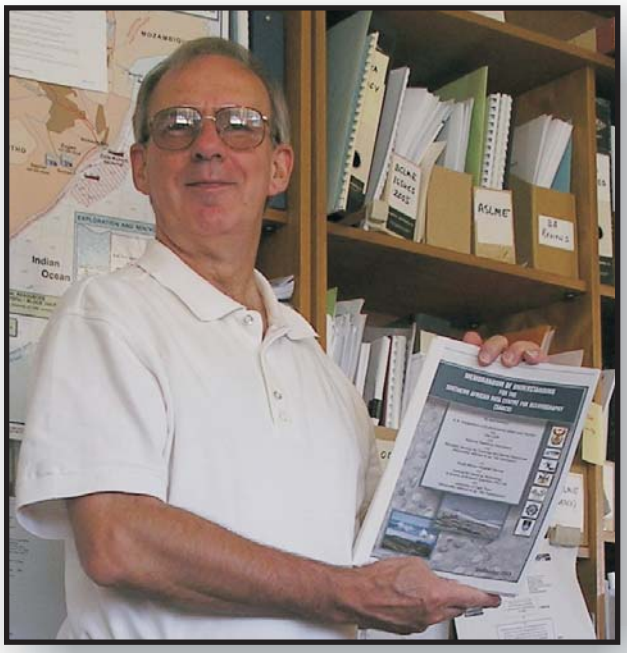
Website: **http://sadco.csir.co.za/**

SADCO is sponsored by ...

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

ISSN 1017-2351

SADCO MoU signed



Marten Grundlingh *holding the finally signed Memorandum of Understanding.*

After approximately two years of circulating from organisation to organisation, the SADCO Memorandum of Understanding, in which several organisations express their support for SADCO, is now finally signed!

Approximately three years ago the SADCO Steering Committee identified the need for the sponsoring members of the Steering Committee, as well as participating members, to express their support for SADCO in a Memorandum of Understanding. This MoU would also serve as a basis for the financial arrangements which are crucial to the successful management and operation of the data centre.

This Memorandum was duly drawn up and circulated for comment, and a year later was ready for approval by the various organisations. It has taken a further two years to obtain the last signature.

The MoU indicates the Objectives and specifications of the data centre, the roles and responsibilities of the National Research Foundation, the database Agency, the Steering Committee and the SADCO manager, financial arrangement, etc.

Each of the sponsoring and participating members will be receiving a copy.

A MILESTONE IN SADCO'S HISTORY: OVERALL DATA QUALITY AUDIT

The first ever, objective assessment of the quality of the data in SADCO, has just been completed, bringing SADCO in line with international quality assessment processes.

The largest part of the previous Newsletter was devoted to the creation of the quality assessment measures in SADCO. Details were provided on the various checks that can be performed on the data, in an attempt to objectively determine (audit) the quality.

Upon completion and validation of the software algorithms it now remained to apply the suite of tests to ALL the profile data in SADCO. This would include data of southern African origin, as well as the data received from various overseas providers [including the data from the World Ocean Database (WOD2005) loaded recently – also reported in the previous Newsletter]. The WOD2005 data had already been screened (and provided with quality flags) by the provider, namely the World Data Centre Oceanography, but these stations were included in the SADCO process so that their results could be included in the overall assessment, and also because some checks (e.g. overland, ship's speed check) are not part of the WDC flagging.

This process has now been completed. The overall summary of the outcome is provided in Table 1.

Statistics on amount of data

SADCO now has data from 5 172 hydrographic surveys (excluding current meter deployments, automatic weather stations, etc). The number is less than the number of surveys in the Inventory, since the latter contains survey information for which the actual data has not been submitted to SADCO yet.

There are 243 566 stations, with measurements at 31 870 895 depths. The global statistics show that temperature is the most common parameter, with 99.8% of the depths having a temperature reading, 41% salinity and 21.6% oxygen. The number of depths with nutrients (including nitrate and phosphate) is significantly lower (< 1%).

The spatial distribution of nitrate and phosphate seem better than their distribution depthwise: 7.3% of stations have at least one nitrate value, and 12.6% of stations have at least one phosphate value.

Examples of quality flags

Detailed description of the quality flags has been provided in the August 2007 issue of the SADCO Newsletter, but brief mention is made below of the type of checks.

a) Not all quality flags are indicative of errors. E.g. a number of stations that have been flagged for failing overland checks, belong to surveys that have been done on the coast, on the beach (e.g. pollution sampling), or in estuaries. Such surveys would understandably receive an "over land" flag.

b) Same as in (a), beach and estuarine sampling may reflect the same date and time since sampling is often done by multiple teams and this may lead to stations receiving a speed flag (the latter is only meant to alert the user of a possible location/time inconsistency). A small number of stations have been flagged for speed because the calculated speed between stations exceeded the threshold of 20 knots. We are aware that some modern ships-of-opportunity that deploy XBTs steam at between 20 and 25 knots, and these would have been flagged. It may be suggested that the threshold of the speed check be increased to 25 knots, to avoid the flagging of data collected from these faster vessels.

A MILESTONE IN SADCO'S HISTORY: OVERALL DATA QUALITY AUDIT (CONTINUED...)

c) 5 276 stations (2.17% of the total) have no "water-physical" records (the latter would include all parameters recorded in the water column, (e.g. temperature, salinity, oxygen, nutrients, trace metals, etc). This would designate stations with data on trace metals located in the sediment, animal tissue or in plankton. Stations with only current measurements would also contribute to this number.

d) The 2262 flags on Depth are due to depth duplicates, reversals in the depths when the instrument yo-yos, or density inversions.

e) Stations with profile flags: When 2 or more observations on a vertical profile fail the annual standard deviation checks, or density inversions, or spikes, the whole profile is flagged as an indication that there may be something amiss with the profile.

f) Observations that failed inversion/gradient checks means that an observation was part of an excessive inversion, or an excessive gradient. E.g. 0.08% of the temperature observations fell into this category. No gradient and inversion checks were applied to oxygen, silicate, chlorophyll and dissolved inorganic carbon (see Table 2, August Newsletter).

g) The envelope check determines whether a particular observation fell within a certain envelope around the profile (information derived from the World Ocean Atlas). E.g. for temperature 1.01% of the observations failed this check, as did 31.19% of silicate observations. Because of this anomalously high percentage for silicate the issue would warrant an investigation. It may be due to a particular set of surveys (from a particular data provider?), or data from areas where the envelopes are not fully appropriate.

h) The broad range check is similar to (g) but the limits are defined for large ocean areas (e.g. the Southern Indian Ocean). Chlorophyll and Dissolved Inorganic Carbon contained higher percentages (8.23% and 5.69%) than the other parameters (which were all below 1%).

I) Very little data contained spikes. This does not necessarily mean that the data was largely spike-free, but that a spike had to be quite large to be flagged.

A table similar to the lower part of Table 1 is now attached to the inventory information of each survey in SADCO, so the user can obtain some immediate insight into the quality of a particular cruise.

The way forward

While the quality (and quantity) audit done on SADCO's hydrographic data holdings is a unique and valuable benchmark, the question arises "So what?"

a) The insight into the quality status has now brought SADCO in line with other global data centres.

b) Should the quality criteria be modified (e.g. the envelopes are "narrowed") the audit can be run again. However, the latter is a time-consuming process to execute for all 240 000+ stations in the data centre.

c) It is planned that raw data extractions will be accompanied by the various flags, allowing the user to decide whether the data is sufficiently accurate for the purpose of the analysis. Experience has shown that some users, investigating processes in data-poor areas, will welcome any data, not just the best quality data. The extraction routines still need to be modified to

incorporate the flags.

d) The introduction of the flagging system allows the database agency the possibility to reduce the (extensive) hands-on quality control of incoming data. Throughout the 1960's, 1970's and 1980's the understanding was that data would be submitted to SADCO only once it has been fully checked and edited. Since 1990 this has proved that data submitted to SADCO contain manifold errors. While SADCO has tried its best to check and, where obvious, correct the data, this is proving an immense task (especially with the discovery of some historic data sets). E.g. for the past 10 years SADCO has inspected each station that reflected a speed check error, and made corrections if the solutions were obvious. This is a time-consuming process. If this had not been done, the 11.03% of data still reflecting speed check errors would have been higher.

e) Should SADCO correct flagged data? In theory, every data centre wishes to contain faultless data, but the cost to improve data quality increases exponentially, and it would be a huge task. The complexity of the issue, and the huge cost involved, probably negates the correction of flagged data. E.g. which parameters should be corrected? Which flags should be tackled? To what level of "accuracy"? Should there be discrimination between instrument types? How should the data provider be consulted? What if new quality algorithms are designed? Should the original and the "corrected" data set be kept? It is obvious that there are big risks involved with such a "clean up" process in that data may be systematically altered in an unnatural way. On the other hand, the global environmental changes are characterised by small variations, calling for data that is accurate to a very high degree. It seems that SADCO's (and other data centres') primary role is to ensure that relevant environmental data IS available and that a first-level assessment is made of the data quality, but that the analysis of the data (including the somewhat subjective "cleaning up") is best left to a person doing this for a particular application.

Acknowledgements

Ursula von St Ange wrote all the software for the QC process, and ran the audit overnight over many weeks. She also compiled the summary statistics (Table 1) and statistics for individual cruises.

We are also grateful for the help of Tim Boyer of NODC (Washington) for clarifying some aspects of the QC algorithms.

Ursula von St Ange

Table 1. Quality audit of SADCO hydrographic data

Surveys

No. of surveys	5172

Stations	Counts	%
No of stations:	243566	100.00%
No of stations failed speed check:	26863	11.03%
No of stations over land:	5406	2.22%
No of stations with no water-physical records:	5276	2.17%

Parameters

Parameter	DPTH	OXY	SAL	TEMP	NO3	PO4	SIO4	CHL	DIC	PH
Total stations with parameter		48363	90098	219625	17862	30738	3026	11191	394	4251
% of Total stations		19.9%	37.0%	90.2%	7.3%	12.6%	1.2%	4.6%	0.2%	1.7%
Total observations	31870895	6877493	13003406	31807494	145994	289984	42031	127106	1776	52328
% of Total depths	100.0%	21.6%	40.8%	99.8%	0.5%	0.9%	0.1%	0.4%	0.0%	0.2%

Flags: Counts

Parameter	DPTH	OXY	SAL	TEMP	NO3	PO4	SIO4	CHL	DIC	PH
No of stns with profile flags		3051	6289	9521	685	2329	241	227	9	35
No of obs failed inversion/gradient checks	2262		76	25083	3496	3				1
No of obs failed envelope check		157179	94435	322823	3682	11937	13111	1531	9	35
No of obs failed broad range check		4858	15260	46905	993	1007	3	10465	101	417
No of obs failed spike check		264	3503	4112	25	26	0	21	9	35

Flags: Percentages

Parameter	DPTH	OXY	SAL	TEMP	NO3	PO4	SIO4	CHL	DIC	PH
No of stns with profile flags		6.31%	6.98%	4.34%	3.83%	7.58%	7.96%	2.03%	2.28%	0.82%
No of obs failed inversion/gradient checks	0.01%		0.00%	0.08%	2.39%	0.00%				0.00%
No of obs failed envelope check		2.29%	0.73%	1.01%	2.52%	4.12%	31.19%	1.20%	0.51%	0.07%
No of obs failed broad range check		0.07%	0.12%	0.15%	0.68%	0.35%	0.01%	8.23%	5.69%	0.80%
No of obs failed spike check		0.00%	0.03%	0.01%	0.02%	0.01%	0.00%	0.02%	0.51%	0.07%

PROGRESS ON SADCO'S WORK LIST FOR 2007/8

Good progress has been made to attend to items on the data centre's work list for 2007/8, with the main items from most of the sectors completed (this includes the two largest items, namely the quality audit and the loading of data from the World Ocean Database.

At the meeting of the SADCO Steering Committee in May 2007 a number of actions were identified that would constitute SADCO's progress on a wide front. These items exclude those that are considered "routine" (e.g. day-to-day management), requests, etc.

The progress with the various tasks is indicated in the attached Table.

In terms of size two tasks are noteworthy:

- Aspects of the Quality Control of the data (see Article this newsletter)
- Loading of data from the World Ocean Database (WOD2005). In the previous Newsletter it has been reported that approximately 36 000 stations had been loaded.

Some foreseen data sets have not been submitted to SADCO yet, and will only be scheduled for processing and loading once they have been received (probably deferred to 2008/9).

SADCO Steering Committee

The November meeting of the Steering Committee is used largely to check on the progress made with the various tasks required by the database agency. On an organisational note, the meeting in Stellenbosch in November 2007 saw some new faces around the table, but was also pleased by the announcement that the chairperson of the Steering Committee, Raymond Roman, was lined up for graduating with his PhD in Oceanography early in December.

Carl Wainman (IMT) and Tracey Gill (SAWS) at the SADCO Steering Committee meeting, November 2007

Sanette Gildenhuys

Cdr Theuns van Niekerk (SAN), standing in for Capt Abri Kampfer (SAN)

Dr Juliet Hermes (Manager, SAEON Egagasini Node), and Kamal Naicker, (Business Manager of SAEON), in a lighter moment.

Table. Status of various items on the SADCO Work list, 2007

Item	Initialised	Underway	Complete	2008/9
Management				
MoU finalise	√	√	√	
SADCO Strategy 2010	√	√		
System				
Software for QC	√	√	√	
Database cleanup	√	√	√	
Loading moored data				
ADCP Namibia	√			√
CMs DeBeers Marine	√	√		
ADCPs MCM*				
Loading profile data				
WOD2005	√	√	√	
Argo float data				√
CTD UCT				√
CTD MCM*				
Loading weather data				
AWS Islands				√
AWS Roman Rock	√	√	√	
AWS MCM*				
Mozambique Wind				√
Port wind data CSIR	√			
Loading other data				
CSIR chemical*				
CLIWOC				√
Wave data CSIR	√			

** No data received*