2nd OceanSITES Data Management meeting

Ref:OceanSITES-DataManagementMeetingReport-01.A

Date 10th -11th April 2008

Location Vienna /Austria

Authors: S Pouliquen, B Burnett

Affiliation : Ifremer, NDBC

Table of Content

Document Change Record

Author	Modification	Issue	Date
S Pouliquen , B Burnett	Initial Version	1A	23rd April 2008
S Pouliquen	Remarks from participants	1B	15th May 2008

The data management meeting was held in Vienna, Austria at 9:00 am on the 10th of April and finished at 1:00 pm on the 11th of April. 31 persons attended the meeting from many countries involved in OceanSites (see attendance list in Steering Committee report). It was chaired by S.Pouliquen/Ifremer and B.Burnett/NDBC.

As part of the introductory words, there was a reminder and general agreement that OceanSITES has to be part of the JCOMM observing system and thus should be coherent and consistent with it. There is a large need and demand for timeseries data and an organized global data system, and the Data Team has to make this happen without re-inventing the wheel.

1. DAC/GDAC STRUCTURE AND IMPLEMENTATION

S. Pouliquen first reminded the participants of the OceanSITES data management structure agreed upon at the last meeting in 2006. The structure relies on three different data management participants (Figure 1):

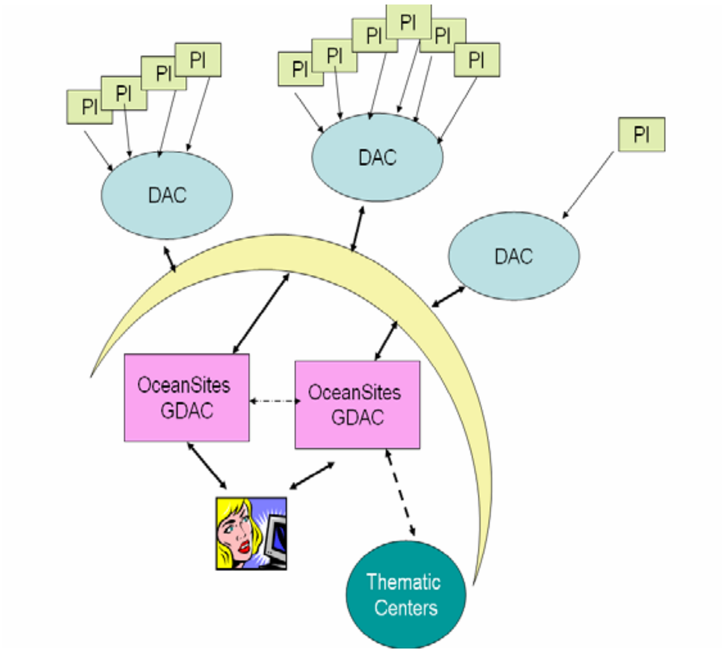

Figure 1 *: Diagram of the OceanSITES architecture*

- **PI (Principle Investigator):** The PI:
    - Maintains specific OceanSITES platforms (one or more sites),
    - Determines what observations are released to the Global Telecommunications System (GTS),
    - Assures that the platform is available and provides reliable information,
    - Provides the DAC with the observations (data) and metadata necessary to serve as an OceanSITES platform, and
    - Quality Controls (QC) post-recovery data according to OceanSITES agreed procedures.

- **DAC (Data Assembly Center):** The DAC:
    - Sets up the OceanSITES "local" server according to the specifications approved by OceanSITES data management group,

- Guarantees data availability from the PI platforms,
- Complies with the agreed upon OceanSITES format,
- Quality Controls real-time data according to the minimum OceanSITES agreed procedures,
- Provides the observations via the Global Telecommunications System (if requested by the PI),
- Provides the data on a FTP server for access by the GDACs
- Organizes the data processing, formatting, data transfer and update with the PIs.
- In the future the DACs would be the place where distributed data may reside

- **GDAC (Global Data Assembly Center)**: The GDAC:
  - Provides a virtual or centralized access to the data that are served by the DACs
  - No data are excluded at the GDAC level, and full high-frequency data sets are available here
  - The GDACs only keep the best version of the data. Additional products like interpolated data are separate optional sets.
  - Check all files daily using the "File Checker" software,
  - Maintains the OceanSITES catalogue, and
  - Synchronizes the catalogues with the second GDAC periodically (at least daily).

Participants at the 2008 meeting agreed to continue with this data management structure, and agreed to Coriolis and NDBC acting as the two GDACs.

Notes:

- One DAC may serve one PI or multiple PIs.
- Thematic centres like CDIAC (for carbon data) can act as DAC for the processing of carbon data for some sites but can also provide all the data for all the PI carbon data providers. This would make the DAC an OceanSITES portal for carbon observations and provide additional scientific validation that may provide enhanced data through the DACs.
- One PI may distribute different data sets to different DACs. For example, meteorological observations may be sent to the National Data Buoy Center while radiation flux measurements might be maintained at another DAC.
- This raises the issue of duplicate datasets residing at different DACs. DACs, overseen by the GDACS, need to carefully aggregate data streams under the same DAC without creating duplicate information in between. There must be a single unique path for each data stream to reach a DAC and GDAC. For each site these streams have to be specified. JCOMMOPS might be able to collect, validate, monitor the data streams to make sure nothing disappears.
- GDACs act as distribution services and will not modify files provided by the DACs.
- The two GDACs will update their data holdings every day by:
  - 1st obtaining the other DACs data index file,
  - 2nd evaluating if the data holdings have changed in the last day
  - 3rd Obtaining the changed observation file by pulling (or pushing) the changed dataset such that both GDACs data holdings are synchronized.
- The GDACs will provide access to OceanSITES observations initially using FTP access to the files provided by the DACs.,
- The GDACs may later provide additional services, like web-based subsetting, merging, etc, and explore serving the data via OPeNDAP servers, or additional services like Open Geospatial Consortium (OGC), Web Mapping Services (WMS) or other types of subscription services.
- Unique WMO ID's have been reserved for OceanSITES (A84nn where A represents the WMO region where the mooring is deployed and nn represents the mooring ID in that region). OceanSITES assumes that there will no more than 99 IDs required for each WMO region. This code is optional and mainly for newly established sites. Sites with long established WMO numbers (e.g., PIRATA and TAO) will not change to the "84" numbers.

1.1. Coriolis GDAC Status

Aurélie Briand presented the status of the development of Coriolis GDAC . Coriolis successfully tested the OceanSITES realtime data flow and implemented daily updates with TAO, KEO and the Animates datasets. A few DACS provided test files to Coriolis, including a TAO CO_2 file from CDIAC, which were checked but mostly did not pass inspection.

Aurélie presented the Coriolis OceanSITES format "file checker" software that is now available for DACs to download and test for their own purposes. The GDACs are expected to test all files with the format checker software whenever a file is submitted. This checker needs two variables to execute (DATA_TYPE = "OceanSITES data" ; - FORMAT_VERSION = "1.0 ") and generates an xml report. During last weeks test, no files passed inspection but a few files only had minor errors. Matthias will provide a sample file that DOES pass the file checker, including CTD data and other data types. There was also a request to implement the possibility to run the checker through a WWW page (Action – Carval / Briand).

As the GDACs provided improved access to OceanSITES datasets, they will continue to improve/standardize the formats used by OceanSITES. Format changes should only occur when necessary, since any changes will be a burden to the DACs and PIs – and all "old" files will need to be reformatted by the DACs and resubmitted to the GDACS. Since there are very few files residing at the GDACs at the present time, everyone in OceanSITES is encouraged to review the format – one more time – and approve the final format (see 2.4).

1.2. National Data Buoy Center (NDBC)

Bill Burnett, NDBC, first discussed the history of NDBC as a long term data provider of marine weather data around the coasts of the United States. However, starting in 2000, NDBC began obtaining and distributing observations from "partners." These partners are designated as U.S. Integrated Ocean Observing System (IOOS) data providers. NDBC receives these marine meteorological, oceanographic (physical) and water quality observations in real-time, quality controls the observations and distributes the data via the Global Telecommunications System (GTS)/web services. NDBC also serves as the Data Assembly Center (DAC) for the Tropical Atmosphere Ocean (TAO) Pacific array and the tsunameter array which covers the Pacific, Atlantic and Gulf of Mexico. NDBC also quality controls and maintains data from 60 oil and gas platforms located in the Gulf of Mexico. Thus, NDBC is well suited to serve as a Global Data Assembly Center (GDAC) for OceanSITES, as well as an OceanSITES DAC.

Next, Bill described how NDBC supports these ~600 platforms by collecting, quality controlling and disseminating the observations in real-time and in delayed mode. Using the NDBC Observing System Monitoring Center (OSMC), OPeNDAP servers and ftp site – NDBC will act as a DAC for physical observations (marine weather and oceanographic – and possibly for biogeochemical variables) for a number of PIs in the United States. NDBC will also serve as the second OceanSITES GDAC and synchronize their OceanSITES files with Coriolis. Bill proposed a form, similar to the form used to maintain the metadata from the 60 oil and gas platforms, to help maintain all the OceanSITES platforms.

Finally, Bill discussed other NDBC collaborations including the U.S. Quality Control of Real-Time Ocean Data (QARTOD), QARTOD to Open Geospatial Consortium (OGC), NOAA IOOS Program Office Sensor Web Enablement/Data Integrated Framework, and the Metadata-Temperature (META-T) Pilot Project with JCOMMOPS.

1.3. Commitment for Setting Dacs

DAC	Site/Station	Real-Time Delayed Mode	Contact point
Bergen University	Station M North Atlantic Seas exchanges S2 Weddell sea	RT/DM DM DM	S Osterhus
PMEL	TAO/Pirata/RAMA Flux reference and enhanced stations, KEO and PAPA	RT/DM	P Freitag
NDBC	TAO WHOI sites SIO sites	RT RT RT/DM	B Burnett
Jamstec(potentially)	TRITON(0-156) JKEO K2 (in future)	DM	M Honda Y Hanafasa
Nioz-DMG	NL-NIOZ data (Irminger Sea, Mozambique channel)	DM	Taco de Bruin
MBARI	OASIS CIMT M0, M1, M2 sites	RT/DM	M MacCann
CCHDO/SIO	CTD data (BATS, HOT, etc)	Semi delayed CTD	S Diggs
WHOI	Stratus/NTAS WHOTS met /flux data	DM DM	N Galbraith
Eurosites	All Eurosites observatoires	RT/DM	M Pagnani
INCOIS	Moored Buoys , current meter mooring array	RT/DM (Data that is being put on GTS)	E Pattabhi Rama Rao
University Hawaii (potentially)	HOT mooring, ship and ancillary data; WHOTS mooring, ocean data	DM (initially) RT(potentially) DM	M Church P Hacker R. Lukas

2. REVISE PRODUCT FORMAT DESCRIPTION

2.1. Feedback from DACs, GDACs, operators

Matthias summarized the format as defined in the OceanSITES user manual version 1.0 and presented some of the inconsistencies with various OceanSITES files that where discovered by the format checker. Matthias noted the deficiencies identified and made a proposal to help improve the file formats. Some of the improvements have already been included in the users manual version 1.1 sent prior to the meeting. The other agreed modifications are included in the list summarized by T Carval (see 2.4).

Matthias agreed to make a proposal to more clearly define which attributes are required for variables, e.g. *.valid min, *.standard name, *.QC indicator, etc.

In order to ease the burden on the users, variables such as LATITUDE and LONGITUDE would be made mandatory but the global attributes - "latitude" and "longitude" - would be removed. However, global attributes such as "southernmost latitude", etc., would remain since they allow for quick automatic screening of disparate datasets. A suggestion was made to similar attributes to conform to a standard since they would also allow for quick automatic screening.. (Action – Carval / Lankhorst)

The OceanSITES filename format was reviewed and is as follows (Users Manual 1.1):
OS_XXX_YYY_ZZZ.nc or:OS_XXX_YYY_ZZZ PARTx.nc
XXX: platform/station code
YYY: Configuration
ZZZ: Variables
PARTx: optional if files to big

The platform/station name should reflect the site name, i.e. a MOVE platform should not be called V407 but MOVE-1, MOVE-2, etc. Configuration can be an identified for the deployment or occupation.

Example : OS_CIS-1_200502_TS.nc: This file contains the T&S data from CIS mooring for the deployment performed in February 2005.

It was agreed that the coherence between the name and the contents should be checked as this will help the GDAC to put the file in the appropriate directory.

2.2. Site and Station catalogue

Aurélie proposed a structure for the Site catalogue. The following structure was agreed (Figure 2.)

- Observatory: Name of the observatory.
- Mooring/platform: Name of the mooring or the platform.
- WMO code: Code assigned to platform if transmitted on GTS.
- Ocean: One of the four oceans where the mooring is located.
- DAC: Name of the DAC. Can be a list if different for Real-Time and Delayed Mode, or per parameters.
- Institute: The institue responsible for the site.
- Contact: Principle Investigator (PI) responsible for the site.
- LAT – LONG: Latitude and Longitude of the area where the mooring is located.
- Start date: Date of the first date available on the GDAC.
- Latest date : Date of the latest date available on the GDAC.
- Update interval: How often the data is updated.
- Information: link to a pdf description of the site provided by the Science team on the OceanSITES site.

site code	platform code	WMO code	DAC	Institute	Contact	southern_latitude	northern_latitude	western_longitude	eastern_longitude	ocean	start date	Latest date	Update Interval	information, web link
CIS	CIS-1	44478	EuroSites	NOCS						Atlantic				Central Irminger Sea, http://….
CIS	CIS-2	44478	EuroSites	NOCS						Atlantic				Central Irminger Sea, http://….
ESTOC	ESTOC-1	13471	EuroSites	NOCS						Atlantic				Estación Europea de Series Temporales del Oceano, Islas Canarias
PAP	PAP-1	62442	EuroSites	NOCS	Lampitt					Atlantic				Porcupine Abyssal

							plain
PAP	PAP-2	62442	EuroSites	NOCS	Lampitt	Atlantic	Porcupine Abyssal plain

Figure 2: Example of Site Catalogue for moorings (NB data are subject to correction)

There was a suggestion to have a data catalogue per observatory while trying to avoid describing the same CTD mooring that might be reporting to different DACS. The draft of the catalogue will be sent to Hester Viola, who will update and maintain the catalogue (in collaboration with others). It was also agreed to improve the index file held by the GDACs by adding start & stop dates, the lat/lon box, etc.. .

2.3. Proposal for Metadata description

John Graybeal submitted a proposal to the group regarding provenance metadata. The following is the proposal:

2.3.1. Recommendation for OceanSITES Data Format Provenance

Introduction. One of the issued raised in the last OceanSITES meetings was recording the provenance of OceanSITES data: where did it come from, and how were the results produced? Such descriptions ideally include not just the data processing events, but the configuration of the measurement systems themselves.

As an outcome of that meeting, John Graybeal and Maureen Pagnani (then known as Maureen Edwards!) were asked to come up with a nominal strategy for documenting the provenance of OceanSITES data files.

Content Standard Options. For sensor systems, there are two comparably mature descriptive languages, SensorML and TransducerML. Of these, SensorML is somewhat more active, comprehensive, and widely adopted in the oceanographic community. Many process description languages exist, including some that can be used to drive processing as well as record it; these are less widely known and less used in the community, and do not describe physical configurations of sensor systems.

Presentation Options. There are essentially 3 presentation options for documenting the provenance: a NetCDF-custom format inside the data file; an external standard file format for describing provenance and other metadata, like SensorML; or adopting part or all of a standard provenance description inside a NetCDF file.

Tradeoffs. Each approach has strengths and weaknesses.

	NetCDF Custom	NetCDF Embedded	External Files
Pros	Stays with file Readily visible (ncdump)	Stays with file Readily visible	Widely interoperable Modular (reusable) metadata descriptions
Cons	Custom format Redundant info in each file	Redundant info in each file Very large, cluttered file Not accessible for other uses	Can get separated from file

While the usual mantra of metadata is to keep metadata with the data file whenever possible, other issues must be considered. In particular, the usability of the metadata descriptions, both by OceanSITES and other purposes, is much enhanced if the descriptions are fully maintained in a standard format like SensorML. Many systems choose to view SensorML description files first, then access NetCDF files using links from the SensorML files. The proposed strategy provides a potential upgrade path for OceanSITES to support more capabilities than NetCDF-based systems can provide.

The modularity of SensorML files comes from their use of xlinking, an XML technology that allows links to other documents to be inserted into XML documents. With this approach, a single SensorML document can describe a sensor system—the platform and sensors that occupy it—while individual documents describe all the individual data products from that platform.

Recommendations. The following recommendations are offered in the order they should be considered. Recommendations 3 and 4 address implementation of recommendations 1, 2, and 5.

1. *Content Standard*. Adopt the SensorML content standard for describing data processing provenance and sensor and platform configurations. Note that SensorML can describe the provenance for each variable in the NetCDF file independently; a best practice would be to use the same variable names in both documents.

2. *Presentation*. Reference the SensorML description files using global variables inside the NetCDF file. Use the global variables that follow for each level of description; note that the same SensorML file could be referenced by several variables:

- Sensor_System_Description: physical configuration of system components
- Data_Processing_Description: processes (including quality control processes) that have been applied to obtain the resulting files

3. *Profile*. Identify the existing best practices of OceanSITES contributors and related activities (n.b.: OOSTethys, BCO-DMO) for describing sensor systems and data processing using SensorML. Compile these practices into a SensorML profile for OceanSITES. This profile should be re-usable by OceanSITES members and other members of the oceanographic community for their own oceanographic data and systems. To maximize participation by the broader community, host development of the profile at the MMI site.

4. *Vocabularies*. For specific terms needed to complete either NetCDF or SensorML metadata fields, create vocabularies using best practices for their development, management, and distribution (n.b. MMI Guides). Managing the vocabularies for openness and visibility maximizes benefit within the OceanSITES community, and awareness, interoperability, and reuse beyond the community.

5. *Quality Assurance*. Evaluate SensorML's applicability to defining quality assurance activities. Identify existing best practices (n.b.: QARTOD) and compile them into a profile that is readily reusable by the larger community. Use the global variable below in the NetCDF file to accommodate a SensorML or other description of QA: Quality_Assurance_Description: describes preparatory processes (tests, cleaning, handling, etc.) to prepare the sensor system(s) for the measurements

6. *NetCDF Metadata Update*. After defining the SensorML profiles, the developers will be in a good position to assess what core provenance information should also be included in NetCDF files, to achieve the best risk mitigation and usability of those files in standalone mode.

References

OOSTethys: http://www.oostethys.org
MMI: http://marinemetadata.org
MMI Guides: http://marinemetadata.org/guides

MMI Vocabulary Guides: http://marinemetadata.org/guides/vocabs
BCO-DMO: http://bco-dmo.org
QARTOD: http://qartod.org
SensorML: http://vast.uah.edu/SensorML
SensorML at OGC: http://www.opengeospatial.org/standards/sensorml

2.3.2. Discussion

Maureen and Thierry agreed to make a proposal adopting SensorML within a month (Action – Pagnani and Carval) . It was pointed out that the OceanSITES data management group should start with a simple and feasible description of the instruments that are on each observatory, similar to the amount of description used by ARGO. Calibration information may be added in the future.

2.4. Agreement on data format and parameter changes

Thierry Carval presented the rational to agree on a revison of the OceanSITES users manual 1.1 format taking into account the feedback from the PIs, DACs and GDACS. The following summarizes the changes:

- Compliance with CF-1.1 released on January 2008: The title, institution, source, history, references and comment global attributes are necessary to follow the CF-1.1 convention,
- Use ISO8601 for dates (strings): yyyy-mm-dd hh:mm:ssZ,
- Remove duplicate information between the global attribute section and the variable section,
- Homogenize use of capital letters,
- Better distinction between nominal locations and observed locations; use a new QC flag for nominal values,
- For moorings without GPS, where there are no time series of time, lat, lon, depth , the user should use a point value for those parameters,
- Enable the possibility of parsing large files (> 100 Mbyes) into pieces,
- Update the parameter table in the users manual and make available on the GDAC,

There was a lot of discussion during the meeting about providing attributes for the parameter's accuracy and uncertainty. The group agreed that parameter accuracy should be mandatory and that providing uncertainty values for observations would be highly desirable.

The parameter table was review and agreed to regarding the following points:

- Standard names will follow the CF convention – however the Data Management Group will work with CF if new standard names are required,
- Only one unit will be permitted by Oceansites even if multiple units are possible.

Parameter Names: In the parameter table a four letter code was defined based on BODC/GF3 standards.

The timestamp should be the center of an averaging/sampling interval, same for depth averages – we will use CF cell methods to describe these in the new format proposal.

Parameter names should always use the same format/name. If different time samplings are required, DACs should continue to use the same parameters – but it was also agreed upon that DACs might need the capability to use several different names . Therefore the DACs can also use a "suffix" appended to the standard parameter name (e.g., TEMP-20 – to describe temperatures collected every twenty minutes).

It the case of interpolated data that being provided with data (i.e., temperature and salinity combined with CO2 measurements) they should be named with a prefix (e.g., CO2_TEMP) so that a user is aware that the observations are not full resolution but averaged for CO2 measurements. It should also use the QC flag located at position 8 (interpolated) and the general attribute for the series shown as interpolated. The group also identified the need to investigate how to record that a measurement was either instantaneous or average. It was agreed to explore the CF convention on this subject.

2.5. Handling OceanSITES observations for model improvement

Operational numerical modeling centers (such as NCEP) are requesting OceanSites data to be provided via the GTS, and that the platforms should be identified as A84xx WMO numbers (This code is optional and mainly for newly established sites. Sites with long established WMO numbers (e.g., PIRATA and TAO) will not change to the "84" numbers) . Numerical modeling centers need to use the data in their forecast and also use them as validation datasets for the reanalysis. The OceanSITES policy states that PIs should be encouraged to distribute their data on the GTS but DACs should have the capability to withhold the data if the PIs do not want the observations released.

3. QUALITY CONTROL

Ideally QC should not be used to fix problems AFTERWARDS, but it QC and QA should be built into the instrument, experiment, sampling design from the start. This requires good instrument design, analytical QC, and conscientious data handling.

QA: standard procedures, training of staff/cross-training, calibration of equipment, lab intercomparisons.
QC: sampling procedures, documentation, storage conditions, analysis with correct primary standards, blanks, review using plots, statistics (but judgement for recognizing anomalous events vs. outliers). It is important to be careful about changing methods, instruments, operators – allow enough overlap/verification time. The cost of very careful manual QC should not be underestimated and sometimes one needs to question whether it is worth the cost.

Data quality control for moored instruments might have three (3) steps depending upon the data received at the DAC.

- Real-Time: A minimum set of automatic procedures to release observations quickly to the GTS or GDAC.
- Near-Real-Time: Telemetered observations are evaluated with past and future data to detect potential drift and/or unrealistic behavior, typical timeframe is 1 month.
- Post-recovery: Post calibrated observations that are released after mooring recovery or after a specified period of time.

3.1. Real-time QC

3.1.1. Proposal for physical/met variables

At the last Data Management Group meeting, a proposal was made to develop QC procedures that could be applied in real time to physical and meteorological observations. An inventory of those procedures was going to be made – unfortunately no inventory was made by the time of this meeting. The Data Management Group encourages each DAC and GDAC to make their Quality Control procedures available to OceanSITES via their public WWW website.

An OceanSITES Data handbook will be written that details the best practices for OceanSITES quality control The document will be developed by: B Burnett/NDBC (lead) , Yasunori Hanafusa /JAMSTEC, N Galbraith/WHOI, M Pagagni/NOCS, T Carval/Coriolis , Pattabhi Rao/INCOIS, and P Frietag/ PMEL

3.1.2. Proposal for biogeochemical variables

Biogeochemical observations will be handled very similar to physical/met data. Real-time procedures will be developed and used for those observations that should be released in real-time, however – some observations will be provided in delayed-mode. Biogeochemical data must include the following information with the observations - date, time, salinity, and temperature – but only those geophysical parameters that are more widely available and which provide a real-time data stream will be targeted initially: Oxygen, nitrate, p(CO2), pH, chlorophyll,…..

The following possible test were identified
- Range checks,
- Continuity and depth checks,
- Correlation with similar observations – if available:

Quality control checks will be developed along with the checks being developed by the Quality Assurance of Real-time Ocean Observations (QARTOD). Any quality control of bio-geochemical observations should be performed at the observation level, and even at the instrument level. Sensors are improving and are able to provide more technical information with the sensor – the data management group should help the DACs understand how this sensor data can help with their overall quality control. The following working group has been developed to lead this effort: A Dickson (lead) , R Lampitt/NOC, F Chavez/Mbari, D Wallace/ IFM-Geomar, T Trull/CSIRO, M Church/Hawaii, R Jonhson/Bermuda, and the manual from this working group should focus on moored autonomous instruments since those are more inhomogeneous and less organized.

3.2. Post-recovery QC

3.2.1. QC approaches for physical/met variables

On post-recovery (the term "delayed mode" was found misleading for mooring data) processing the data management group identified the need to collect and assemble best practices including minimum standard tests , sensor calibration, clock drifts, maybe attaching of instruments to a CTD for calibration casts, and comparison with other data collected via delayed-mode. The group also agreed that it was important to solicit and receive feedback from all users. The QC handbook should also draw on existing documents from e.g. WOCE current meter DAC. It will be a common document since the handbook will address both real-time, near real-time, and delayed mode QC procedures.

3.2.2. QC approaches for biogeochemical variables

Mathiew Church presented what is done at the HOTs station (a CTD occupied site). He presented QC procedures for observations collected during cruises, processed in the lab and at on-sea moorings. All these procedures are encouraged to be used by the DACs.

3.2.3. Other quantities Geophysical, acoustic, image data, etc

John Orcutt presented what is done at geophysical stations. He presented QC procedures for observations collected during cruises, processed in the lab and at on-sea moorings. All these procedures are encouraged to be used by the DACs. The OOI CI might take in existing OceanSITES data to practise/test the system. The main geophysical data that can practically be part of OceanSITES are the DART data. We may be able to share logistics to deploy/co-locate OBS and OceanSITES data. If an OceanSITES station provides seafloor telemetry capability, some geophysical sensors might be able to exploit those data channels.

4. DATA POLICY

Following a request from the CLIVAR group, the data management group agreed to endorse the CLIVAR data policy and accepted a page that Uwe prepared for the specificity of OceanSITES (see Annex 1).

5. COMMITMENTS TO MAKE THE DATA FLOW

All present operators of sites and representatives of DACs agreed to make their data available in the new format, follow the agreed procedures, and to make the data flow, both in real-time and in post-recovery mode.

6. WRAP-UP AND ACTIONS FOR THE DATA MANAGEMENT PART

1. All operators and DAC's will start to format and send data.

2. Update the data flow documentation and provide guidelines for DACs on how to submit data to be included in the GDAC data Handbook (S. Pouliquen).
3. Update the Users Manual to reflect the format changes agreed upon at the meeting (T. Carval).
4. Work with Science Committee to update the OceanSITES catalogue (H. Viola and B. Burnett).
5. Provide SensorML format for mooring description to initiate discussion (T. Carval / M. Pagnani).
6. Propose new form to survey/declare official OceanSITES platforms (B. Burnett).
7. Coordinate GDAC Data Handbook and submit draft to Data Management Group (B. Burnett)
8. Propose Quality Control best practice for physical and met parameters (B. Burnett lead)
9. Propose Quality Control best practice for Bio-Geochemical parameters (A. Dickson)

7. ANNEX 1 CLIVAR RELATED DOCUMENTS

CLIVAR Data Policy

Introduction

CLIVAR, a global multidisciplinary climate research project, requires a wide range of data and needs data centres to ingest, quality control, archive and distribute these data. The CLIVAR data policy provides guidelines for how these data should be handled in a consistent manner so as to achieve the project's scientific objectives. The policy aims to strike a balance between the rights of investigators and the need for widespread access through the free and unrestricted sharing and exchange of CLIVAR data and metadata. CLIVAR data policy is intended to be fully compatible with IOC [1], WMO [2], GCOS [3] and GEOSS [4] data principles.

The multidisciplinary nature of CLIVAR and its subprogrammes means that the principles enshrined in the Data Policy must be applied to data in each subprogramme's implementation plan

Definitions used in the CLIVAR Data Policy

1. CLIVAR data

"CLIVAR Data" consists of directly observed data, derived data, gridded fields, and other data products generated and/or used within CLIVAR to further its scientific goals. CLIVAR data and related products can be categorized in terms of those that are specifically sponsored or endorsed by the international CLIVAR programme, those generated by other related bodies and programmes such as the World Weather Watch of the WMO, GCOS, JCOMM, and other projects of the WCRP and those generated by relevant national and institutional CLIVAR-related projects and programmes. CLIVAR should strive to ensure that all data relevant to CLIVAR are accessible freely and without restriction, including those collected by other projects and programmes.

2. Metadata

Metadata is defined as the descriptive information such as content, quality, condition that characterizes a set of measurements.

CLIVAR Data Policy and Principles

For CLIVAR to succeed, high-quality data and metadata need to be collected, processed and exchanged without significant delay in a free and unrestricted manner. This was recognized in the CLIVAR Implementation Plan [5] in discussing 'The Principles for CLIVAR Data'. CLIVAR data policy is enshrined in the CLIVAR data principles below:

1. Free and unrestricted exchange

All CLIVAR data should be made available freely and without restriction. "Freely" means at no more than the cost of reproduction and delivery, without charge for the data itself. "Without restriction" means without discrimination against, for example, individuals, research groups, or nationality. In exceptional circumstances involving highly specialized or experimental data, principal investigators may temporarily limit access until such time as the data can be adequately validated.

2. Timely exchange
CLIVAR investigators should make data available voluntarily and with minimal delay, preferably also in real-time, to maximize their value to CLIVAR. In cases where extensive post-processing of delayed mode data is needed before a final research quality data set can be generated, early release of a preliminary version of the data is required.

3. Quality control
CLIVAR investigators retain the primary responsibility for the quality of the data they produce and distribute. Data originators and those generating climate data products are required to ensure that their data meet international quality standards wherever possible.

4. Metadata
Metadata are required to enable the use of data without ambiguity or uncertainty. Metadata for CLIVAR data sets will be developed and managed in accordance with international standards.

5. Preservation of data
Long-term survival, integrity, and access to CLIVAR data must be preserved for future generations. Internationally agreed standards should be used for the acquisition, processing, and final archival of data and metadata. Data distributed in real and near-real time should, wherever possible, be replaced in a delayed mode after it has undergone quality control and full documentation.

6. Plan for reuse in reanalysis
While datasets will be used individually and in combination for research purposes, the sum total of all CLIVAR and CLIVAR-relevant data will have great value in reanalysis activities. To aid this, uniformity of data format and quality should be a high priority.

7. Easy access
CLIVAR encourages the use of the most recent advances in communication to ensure widespread access to data collected under auspices of the programme.

8. Use of existing national and international mechanisms and centres
Where feasible, CLIVAR will use existing national and international mechanisms for the exchange and storage of oceanic and atmospheric data, and build on the data management structure of existing programmes. In this way, the effectiveness of the data system will be improved by reducing redundancy and duplication and identifying opportunities and system economies, with financial costs minimized.

9. Reporting Requirements
Data and metadata should be submitted to recognized data assembly centers as well as to appropriate national and international archival institutions so that the collected information may be safeguarded for future analysis. Inventories of data and related information should be readily accessible and updated as needed on a routine basis.

Observing System Data Access: OceanSITES Timeseries System

Prepared by the OceanSITES Steering Team
Version 1 April 2008

1. Program Definition

The OceanSITES timeseries program is the global network of open-ocean sustained timeseries sites being implemented by an international partnership of more than XXX nations. OceanSITES provides fixed-point timeseries of various physical, biogeochemical, and atmospheric variables at different locations around the globe, from the atmosphere and sea surface to the seafloor. OceanSITES' objective is to build and maintain a multidisciplinary global network for a broad range of research and operational applications including climate, carbon, and ecosystem variability and forecasting and ocean state validation. All OceanSITES data are publicly available via the internet and partly via the GTS. Information on the OceanSITES program, the location and architecture of sites, and access to OceanSITES data can be obtained at http://www.OceanSITES.org .

2. Data Flow and Quality Control

OceanSITES data are partially transmitted in real-time and relayed to regional or national Data Assembly Centers (DACs). The DACS are responsible for applying automated real-time quality control tests to identify and flag grossly bad data. Data that pass the automated QC tests are broadcast on the GTS, apart from those purposely withheld for validation purposes. All data, with flags, are relayed to Global Data Assembly Centers (GDACs) in Brest, France, and at NDBC/USA. The GDACs maintain complete (mirror) datasets, and make all data available from one place in a unified format, initially via ftp directories, later through user-friendly interfaces.

Delayed-mode OceanSITES data are also available, having been post-calibrated and visually inspected by a principal investigator or comparable expert. Delayed-mode data files include both the original measurements and the best available corrected measurements as determined and flagged by the expert.

Documentation of automated real-time quality control procedures will be available through the web portal (www.OceanSITES.org).

3. Timeliness of Distribution

Realtime OceanSITES data that pass the gross real-time QC tests and that are not meant to be withheld, are distributed via the GTS, within 24 hours for about 90% of data. Effort is ongoing to improve timeliness. All real-time data are flagged and available from the GDACS within a few hours of the GTS distribution.

The timeliness objective for delayed-mode distribution is for the research-quality data to be available within 12 months of instrument retrieval. The sensors require post-calibrations and time is needed for the expert inspection of timeseries from a given instrument.