3rd OceanSITES Data Management Team meeting

Date: 18th September 2009

Location: Venice, Italy

Authors: Bill Burnett (NDBC)

Meeting information: http://www.jcomm.info/oceansites2009

Table of Contents :

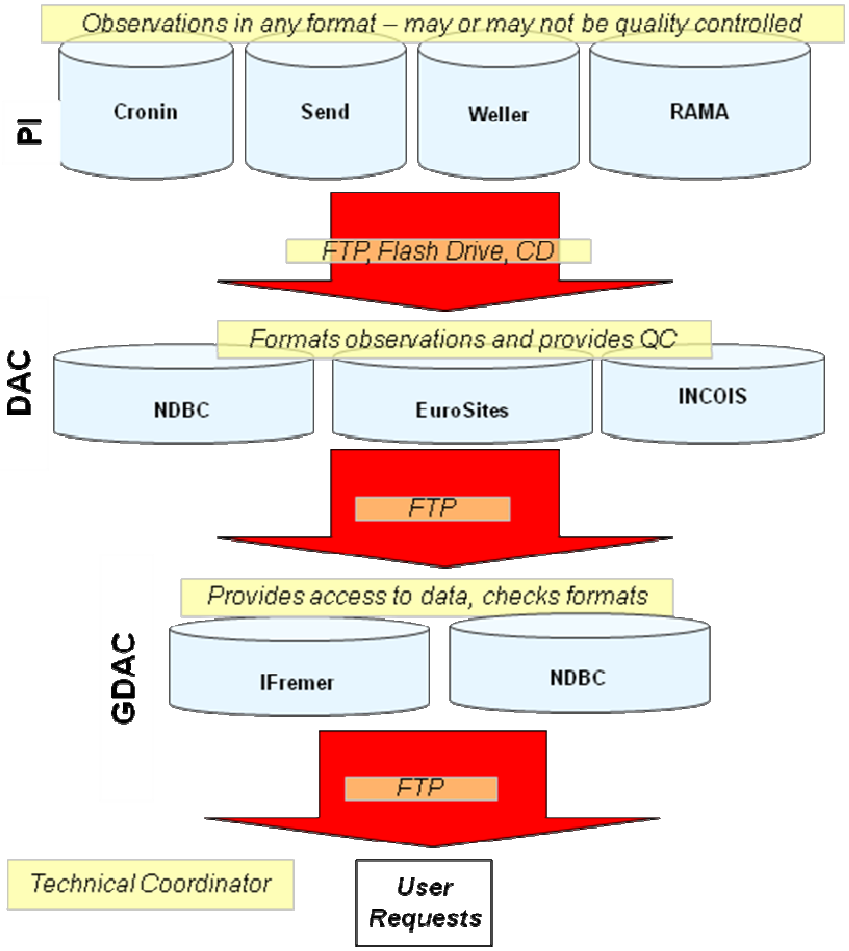
Attendees:

Taco de Bruin, Bill Burnett, Thierry Carval, Meghan Cronin, Steve Diggs, Nan Galbraith, Makio Honda, Alex Kozyr, Richard Lampitt, Matthias Lankhorst, Mike McCann, Chris Meinen, Maureen Pagnani, Sylvie Pouliquen, Pattahbi Rao, Cecile Robin, Eric Schulz, Uwe Send, Derrick Snowden, Hester Viola, Robert Weller, Scott Woodruff, John Orcutt

For more detail about attendees, see the report of the 7th OceanSITES Steering Team meeting.

1. Review of OceanSITES Data Holdings

Bill Burnett presented an overview of the current OceanSITES data management architecture, describing how the data flow proceeds from the Principle Investigator (PI) to the Data Assembly Center (DAC) to Global Data Assembly Centers (GDACS). He also reminded the participants about their roles and responsibilities.



He stressed that the GDACs are the ultimate destination (pre-archival) for observations – ensuring ease of access for any users of OceanSITES data. The GDACS also ensure the validity of the data holdings. However the role of quality control and formatting of the data resides at the DAC level. Even with these roles, the GDACs can, and should, play a role in helping DACs set up their data delivery systems to the GDAC. PIs need to be aware of all the metadata requirements and be able to provide metadata to the DAC once per deployment. This supports the requirement for the DAC to get the metadata before the actual data.

The following list outlines the eleven (11) DACs "validated" as operational DACs during the meeting with the appropriate points of contact (this list needs to be filled out with the POCs).

- NDBC (Point of Contact – Bill Burnett)
- EuroSITES (managed by NOC and IFREMER)
- INCOIS
- JAMSTEC
- MBARI
- WHOI – Nan Galbraith
- CCHDO
- NIOZ
- PMEL
- IMOS
- CDIAC

Bill Burnett showed the two GDAC FTP sites to the participants. He compared and contrasted the GDAC sites and data holdings. Presently the NDBC GDAC ftp directories are organized by "DAC" while the IFREMER GDAC is organized by "Observing Networks and Sites Organizations." The team resolved to organize OceanSITES data holdings by Observing Networks and Sites Organizations. Sites will be defined in the Sites Catalogue managed by the Technical Coordinator at JCOMMOPS and described in the User handbook.

The group noted that the definitions for Observatory, Network, Site, Platform, deployment should be finalized and that these "sites" should be registered at the Project Office. In order to provide an easy and clear access to the different mooring deployment descriptions and data it was agreed to organize the GDAC sites with the following directory tree : Site/ Platform/ deployment . The deployment directory will contain the metadata file (presently the Doc file, in the future sensorML file or directory) and the different data files. This will ensure that the GDAC ftp sites will be properly maintained and mirrored. Files should also be listed as either real-time or delayed mode. One issue that must be resolved is how to properly store real-time and delayed-mode versions of the same data on the GDAC servers. The GDACs will continue to hold the "best version" but more discussions with the Steering Team will be required to understand their requirements and GDAC capabilities.

2. Priorities for expanding OceanSITES Data Holdings

Hester Viola (JCOMMOPS Technical Coordinator (TO)) presented the Site catalogue built according to last meeting's specifications. The Team agreed that the "Observatory-Network" field was a subjective field that may not be reliable. What will be relied upon is the "Site" field. The geographic meaning of Observatory is still interesting but will be used as a comment in the site catalogue. The TC will revisit the "SITE" field names with the eleven DAC managers (once defined) to ensure the fields are filled out correctly. Participants suggested that the TC should add additional columns to the catalogue to describe the purpose of the platform (flux measurements, carbon, etc…) which is available in the other excel sheet (i.e., the sheet which describes all the sites that will potentially distribute data). The spreadsheets will be moved from Excel files to a database once the fields are finalized. One participant pointed out that it would be useful to know if a site is operational, acquiring data, but is not yet providing the data to a DAC. This can be done at JCOMMOPS through the catalogue and the index files available at GDAC. It would also be useful to plot each mooring's

time scale for the period covered on the OceanSITES website – however quite a bit of time and effort will be required by the TC and the DAC managers to develop this plot.

Another issue is the use of interpolated data for OceanSITES. The group indicated that any interpolated data should be labeled in a different file. Format and head issues will need to be resolved in further discussions.

Data providers then provided a list of current and future commitments for their sites:

- Norway : 3 sites are identified but only one is in EuroSITES . No data yet
- PMEL : 4 multidisciplinary TAO moorings provided through PMEL. The team recommended that NDBC include all TAO data in OceanSITES. Observations from the PIRATA and RAMA arrays will need to be provided by PMEL.
- MOVE : no real-time data but will send 10 years of data to GDAC . One real-time mooring will be deployed soon
- WHOI : doing both real-time and delayed-mode directly. First data sent (NTAS, WHOTS, STATUS) to IFREMER.
- JAMSTEC : willing to provide Triton and JKEO but not ready yet
- NIOZ not submitted yet
- MBARI : Data are on MBARI server . Should have been collected by GDACs. NDBC is working with MBARI to resolve the issue.
- CCHDO : Not delivered yet (HOTS , BATS) CTD data are at CCHDO and finalizing checking with the format Then will be working on bottle data
- INCOIS : 3 sites provided with metadata
- EUROSITES: Maureen is working on providing the data. She is facing the problem that one institute is asking for the ability to trace users who download the data. The Data Team agreed that users will not be asked to provide information in order to use the data – everything will remain free and open.

Participants were reminded that once the data flow is prepared for the GDAC, they should work on providing their past (historical) data to OceanSITES as feasible since long time-series of observations are important to OceanSITES.

3. Review of OceanSITES Data Holdings

Bill Burnett discussed the metadata spreadsheet which was developed during the past year. As an initial step towards collecting significant metadata about OceanSITES platforms, NDBC and the TC developed a Word document which was submitted to some of the PIs and DACs last year. The

document requests PIs to submit significant metadata about a site. Depending upon how many instruments are on a given OceanSITES platform, the document could be very large. One participant pointed out that all of the metadata she had were already in the NetCDF file and that filling the Word file takes twice the effort. She advocated for submitting the data in NetCDF to avoid duplication. Another participant pointed out that NetCDF files are not robust enough to provide significant metadata (i.e., calibration information, deployment photos, etc…). Therefore that participant felt that SensorML or a similar format should be developed to enable the providers to provide more structured descriptions of the platforms.

There was a consensus to generate metadata information in separate files that will allow interoperability with international catalogues To ease integration with international data exchange networks, GDACs have, or are setting up, OPeNDAP servers that will enable additional services at the each GDAC and provide robust dissemination services apart from FTP. Similar servers could be set up at DACs allowing more frequent updates of OceanSITES data. The next step would be to implement Sensor Observation Services (SOS) servers that would allow interoperability with the rest of the world. NDBC is operating a SOS server at http://sdf.ndbc.noaa.gov. This work was funded by the NOAA IOOS Program Office as a part of their Data Integrated Framework. NDBC also agreed to set up RSS services similar to the one they provide for their NDBC observations.

The question of timeliness of data was raised and it was agreed that the requirement is still to have a 24 hour delay for real-time and 12 months for delayed mode.

Finally, authors of journal papers using OceanSITES data should be encouraged to cite the relevant OceanSITES PIs. Therefore suggested citation requests for users of OceanSITES data will be incorporated into the NetCDF files, and will also be posted on the OceanSITES website. Furthermore, the GDACs will monitor statistics from OceanSITES servers to provide metrics on the utilization of OceanSITES data.

4. Format Issues

4.1 Metadata Issues

Last year the data management group recommended that Site/Platform/deployments should be described more thoroughly with additional metadata information. Bill and Sylvie proposed moving all metadata sheets to SensorML in order to describe a site, a platform and a deployment. This will allow OceanSITES to maintain metadata more properly, and associate OceanSITES data with the observations. It will also allow OceanSITES to collect more information about the stations. Any plans to implement SensorML will be at least a year away; however NDBC noted that they are implementing SensorML for their information through the Data Integrated Framework.

Mike McCann presented a brief on MBARI's metadata system and how the metadata is used. He stressed the importance of collecting metadata in the early stage of a deployment. In a metadata system, information about the data processing system is stored on a server. There is a need for a standard "serialized" representation that would allow hierarchical description of a deployment. SensorML might be the optimal way to obtain and disseminate the information. There is also a need for a Java GUI that would guide the users to fill the SensorML data. One participant pointed out that

the tracking of data processing information has not been addressed properly in OceanSITES, and it will need to be addressed soon.

Cecile Robin, from IFREMER, presented a SensorML description of a deployment that would replace the Word document ("Sheet") presently used by some to document the sites. SensorML V1 has been adopted as an OGC standard. In her proposal, she presently describes a platform as composed of instruments. One participant pointed out that she should keep the flexibility to describe a system to be composed of systems - hierarchically in as many sublevels as requested. This description would allow PIs to indicate that a platform is acquiring more parameters than the ones available in NetCDF. Clarification of what is a component needs to be provided (one example – should the group capture information about the datalogger that was used during the deployment?). The Team agreed that there are links between NetCDF data files and metadata description, and that there should be at minimum one metadata file per data file. A deployment file might be split into different files (example : Surface met parameters and ocean parameters are in two different files) . The Team should also study how to create metadata files that will link the different parts of a deployment, and also describe a site as a suite of deployments.

5. Update of GDAC Management Plan and User Handbook

5.1 Review Format Issues

Mathias Lankhorst summarized the different format issues that were raised during the past year. He addressed the following points:

- How to attribute pressure measurements made along a mooring chain. The group agreed to add two attributes, "sensor_mount" and "sensor_orientation" to the data file, and also provide a list of authorized values. The list also needs to be validated together with the default values.

- Do we keep the raw and adjusted best value? The group decided not to implement two types of files – raw and adjusted – due to the difficulty in maintaining the file sets. However, the group did highlight the need for an archive repository, and requested information from the U.S. National Oceanographic Data Center (NODC) to determine if they would serve as a long-term repository for OceanSITES.

- What should the group do with bad values in the dataset? The default solution is to keep the data and flag it appropriately. But it is also up to the PI to correct the dataset and indicate that the observation was modified with the appropriate flag. Obviously this issue needs to be discussed and resolved.

- PIs should include the sensor serial number and sensor name (i.e., manufacture name) in the metadata document sheet until the group determines how to properly code the data in the NetCDF file. Modify the manual to indicate that DEPTH is required.

- Latitude/Longitude/depth should be able to vary with time : check the manual

- File name conventions OS_XXX_YYY_ZZZ_Partx.nc

The ZZZ group is supposed to indicate the list of parameters on the platform – however the group is probably not adequate since platforms might measure multiple parameters, making the string very long. The group proposed to add the information to an index file which contains the list of all parameters measured by the OceanSITES platform. The problem is that users cannot look at the file name and instantly understand what type of information is contained in the file.

The file name convention would be changed as follow OS_XXX_YYY_T_Z_Partx.nc

T : R for RT, D for DM, P for post recovery, M for mixed (contains both real-time and delayed-mode data)

Z: The group will make a proposal to indicate in the name which kind of data it contains

1. The GDACs should develop a mechanism to prevent two DACs from submitting different datasets with the same name.

5.2 Parameter list in users manual

- The Team should update the parameter list in the user manual to reflect new parameters that are in the NetCDF file but not in the data file.
- The Team should specify which Temperature and Salinity scales are used: i.e., the reference scale specifying ITS_90 or PSS-78.
- In the user manual there is a Parameter Name and Standard Name Table. The Team should develop a mechanism to enable the DACS to update this list when required. In particular there is a need for additional names for Carbon in CF.
- Modifications to the Index File – the group should add a parameter list to the file and add lines to indicate when the files are updated.

Bill Burnett developed the first draft of the GDAC Data Management Plan but indicated that further edits and reviews are required. Bill will try to incorporate some of the information in the NDBC Quality Control handbook into the OceanSITES handbook. Further additions may be made by the Quality Assurance of Real-Time Ocean Data (QARTOD) meetings which are held in the United States every year. This should address over 80% of the OceanSITES variables. The missing variables will be identified and submitted to the co-chairs for review and action. Thierry will circulate an updated version of the User Manual taking into account this meeting discussion.

6. Next meeting

There will a Data Management meeting in Paris around late March, 2010 in conjunction with the IMDIS meeting. The Data management team will also hold WebEx meetings to address the numerous data management issues brought up during the meeting. These meetings will require representatives from each Data Assembly Center. The Team feels it is important to have clearer guidance from the Steering Team which means that some representatives should attend the meeting.

7. Action Items or Questions to be addressed:

1. Each institute named in this list should provide the data management team with the name of a representative who will be a part of the data management mailing list and represent the institutes in the OceanSITES data management meetings. This "Point of Contact" will also be notified when data holdings have not been updated after an appropriate level of time. Action - DACs

2. Does the Institute of Marine Research (IMR) want to serve as a DAC or rely on the EuroSITES DAC to distribute their data to OceanSites Action - IMR

3. The group decided that the GDAC ftp sites should be arranged by site organization (i.e., similar to IFREMER's site). Action – NDBC.

4. Definitions for Observatory, Network, Site, Platform, deployment should be finalized as soon as possible and that these "sites" should be registered at the Project Office. Action – Technical Coordinator with the help of the Data management team.

5. The Project Office Coordinator will revisit the "SITE" field names with the eleven DAC managers (once defined) to ensure the fields are filled out correctly. Action – Project office

6. Post citation requests for users using the citation provided into the Netcdf files on the OceanSITES website. Action – Steering Team chairs

7. Cecile to work with Nan and Mike to describe more complicated site than ESTOC in SensorML. Action - Cecile

8. IFREMER and NDBC will compare their implementation of SensorML and work for a proposal for OceanSites in 2010. Action – Cecile and Burnett

9. Request information from the U.S. NODC to see if they would serve as a long-term archive repository for OceanSITES. Action – Bill Burnett

10. The group will make a proposal to indicate in the name which kind of data it contains. Action – Burnett

11. GDACs develop a method to ensure two DACs don't submit different datasets with the same header. Action – Carval and Burnett

12. The Team should update the parameter list in the user manual to reflect new parameters that are in the NetCDF file but not in the data file. Action – Burnett

13. The Team should specify which Temperature and Salinity scales are used: i.e., the reference scale is either ITS_90 or PSS-78. Action – Burnett

14. In the user manual there is a Parameter Name and Standard Name table. The Team should develop a mechanism to enable the DACS to update this list when required. Action – Burnett.

15. In particular there is a need for additional names for Carbon in CF. The CF Standard Names Committee has added several new terms for carbon and related studies. A PI or DAC that requires new terms to describe any variable should review the CF discussion list first before negotiating a new term. Action – PI or DAC

16. Modifications to the Index File – add a parameter list to the file and lines for when the files are updated. Action - Carval to update the manual-GDACS to implement

17. GDACs produce regular statistics on data access

18. GDACs implement a RSS mechanism to inform users on data updates

19. GDACs implement an OpenDAP access on top of the FTP server

For more detail about action items, see the report of the 7th OceanSITES Steering Team meeting.