GROOM

Gliders for Research, Ocean Observation and Management

FP7-Infra-2011-2.1.1 "Design Studies"

Deliverable D3.3

Report on Global Data Centre organization for Gliders

Due date of deliverable: 31/12/2013

Actual submission date: 20/12/2013

Partner responsible: Ifremer

Classification: PU
Grant Agreement Number: 284321

Contract Start Date: October 1st, 2011
Duration: 36 Months

Project Coordinator: UPMC
Partners: UPMC, OC-UCY, GEOMAR, HZG, AWI, UT, FMI, CNRS, IFREMER, HCMR, CMRE, OGS, UIB, NERSC, CSIC, PLOCAN, SAMS, UEA, NERC.

Project website address http://www.groom-fp7.eu

Table of contents

I. CONTRIBUTORS AND ACKNOWLEDGMENTS

This report was prepared as part of Task 3.2 (Data flow and processing) of Work Package 3 (Scientific Innovation) of GROOM.

S Pouliquen and T Carval of the IFREMER/France, were the lead authors and editors of the report, with significant assistance from following persons that worked together for a week in December 2012 to consolidate the data system and prepare common tools to process the glider data

Thierry CARVAL	Ifremer	France
Jean-Philippe Rannou	Altran	France
Justin BUCK	BODC/NERC	UK
Mark Hebden	BODC/NERC	UK
Lise Quesnel	BODC/NERC	UK
Bartolomé GARAU	CMRE	Italy
Daniele CECCHI	CMRE	Italy

This report also relies on input provided by the WP3 partners who participated to the two data management workshops that were held in Paris in October 2012 and in Trieste in June 2013.

The list of participants to these workshops was:

Nadin Ramirez	UDEC	Chile
Angelos Hannides	OC-UCY	Cyprus
Chrysostomos ELEFTHERIOU	OC-UCY	Cyprus
Dan HANES	OC-UCY	Cyprus
Elodie GODINHO	DT-INSU	France
Karim BERNARDET	DT-INSU	France
Laurent BEGUERY	DT-INSU	France
Loïc PETIT DE LA VILLE	Ifremer	France
Sylvie POULIQUEN	Ifremer	France
Thierry CARVAL	Ifremer	France
Antony BOSSE	LOCEAN	France
Laurent MORTIER	LOCEAN	France
Pierre CAUCHY	LOCEAN	France
Pierre TESTOR	LOCEAN	France
Vincent TAILANDIER	LOV	France
Alice PIETRI	GEOMAR	Germany
Gerd KRAHMANN	GEOMAR	Germany
Travi LIBLIK	GEOMAR	Germany
Gisberg BREITBACH	HZG	Germany
Lucas MERCKELBACH	HZG	Germany
Dimitris KASSIS	HCMR	Greece
Leonidas Perivoliotis	HCMR	Greece
Bartolomé GARAU	CMRE	Italy
Daniele CECCHI	CMRE	Italy
Reiner Onken	CMRE	Italy
Antonio BUSSANI	OGS	Italy
Elena Mauri	OGS	Italy
Giulio Notarstefano	OGS	Italy
Riccardo Gerin	OGS	Italy
Erik BRUWIK	UiB	Norway
Svein ØSTERHUS	UiB	Norway
Agnieszka BESZCZ	IOPAS	Poland

Simon Ruiz	CSIC	Spain
Alvaro LORENZO	PLOCAN	Spain
Carlos BARRERA	PLOCAN	Spain
Emma Heslop	SOCIB/CSIC	Spain
Joan Pau BELTRAN	SOCIB/CSIC	Spain
Justin BUCK	BODC	UK
Mark HEBDEN	BODC	UK
Estelle DUMONT	SAMS	UK
Toby SHERWIN	SAMS	UK
Bastien QUESTE	UEA	UK
Jan Kaiser	UEA	UK
Sunke SCHMIDTKO	UEA	UK

II. OBJECTIVES

The objective of this report is to propose a data management system for European Gliders within GROOM project that could be sustained after GROOM and also be extended to other countries within an international glider program. Therefore such system has to be developed in coherency with what is under development at international level for gliders (IMOS in Australia, IOOS in USA...) and interoperable with other JCOMM networks (Argo, OceanSITES …) or IODE standards (SeaDataNet/EU, QUARTOD/USA, GOSUD, GTSPP…)

Main goals of an enhanced data management system for gliders in Europe are to:

- Provide a unique access point to the European gliders
- Improve the data coherency in term of format, quality, processing chain (clearly documented)
- Set up a capability able to serve both operational (within a few hours) and research (best quality after calibration and validation) users

At the start of GROOM project glider data are managed by the different research communities using their own methods and data exchange in real time is working on a best effort schema through EGO but without any commitment nor from providers, nor from data managers. This allows Coriolis to provide glider data as profiles to the Copernicus Marine Core Service operated by MyOcean & MyOcean2 FP7 projects using some agreed processing methods developed within FP5-MFSTEP, FP6-MERSEA and FP7_MyOcean projects. Presently there is no agreement neither on Real Time Quality Control procedures (RTQC) or Delayed Mode QC (DMQC) but best practices on RTQC have been developed through MyOcean In-Situ Thematic Assemble Centre (INSTAC) but they are not widely known by the glider teams.

At the start of the GROOM project the situation was pretty heterogeneous:

- Some institutes activities covers the whole data management spectra from piloting their glider to data processing the data in real-time and delayed mode.
- Some institutes are only interested in real time data or in delayed mode one. In the latter case the real-time data stream is only used for piloting the glider.
- Some countries already started to organize the glider data management activities and data provision to operational and research communities.
- Most partners provide their data to LOCEAN to be visible on EGO WWW site.
- A certain number of partners provide their real-time data to Coriolis for provision to operational users.

III. SOME DEFINITIONS

A glider is moving platform that is steerable. It can have a propeller and this information must be recorded in the metadata. The data can be provided in Near Real Time which means within a few hours from acquisition.

Glider data goes through different levels of processing:

- Level0 : Data provided by the glider without any unit transformation or geophysical interpretation.
- Level1: Geophysical parameters with a quality indicator set up by automatic QC procedures together with the data acquired by the glider. This is the level shared in Near Real Time.
- Level2 : Geophysical parameters calibrated after glider recovery together with quality flag information, if possible error estimation together with the non-corrected data provided at Level1. This is the level shared in Delayed Mode

- Level3 and after : Product derived from glider data (gridded fields, additional parameters calculated …) This is not addressed in the present GROOM data management activities.

IV. ACTORS AND DATA FLOW

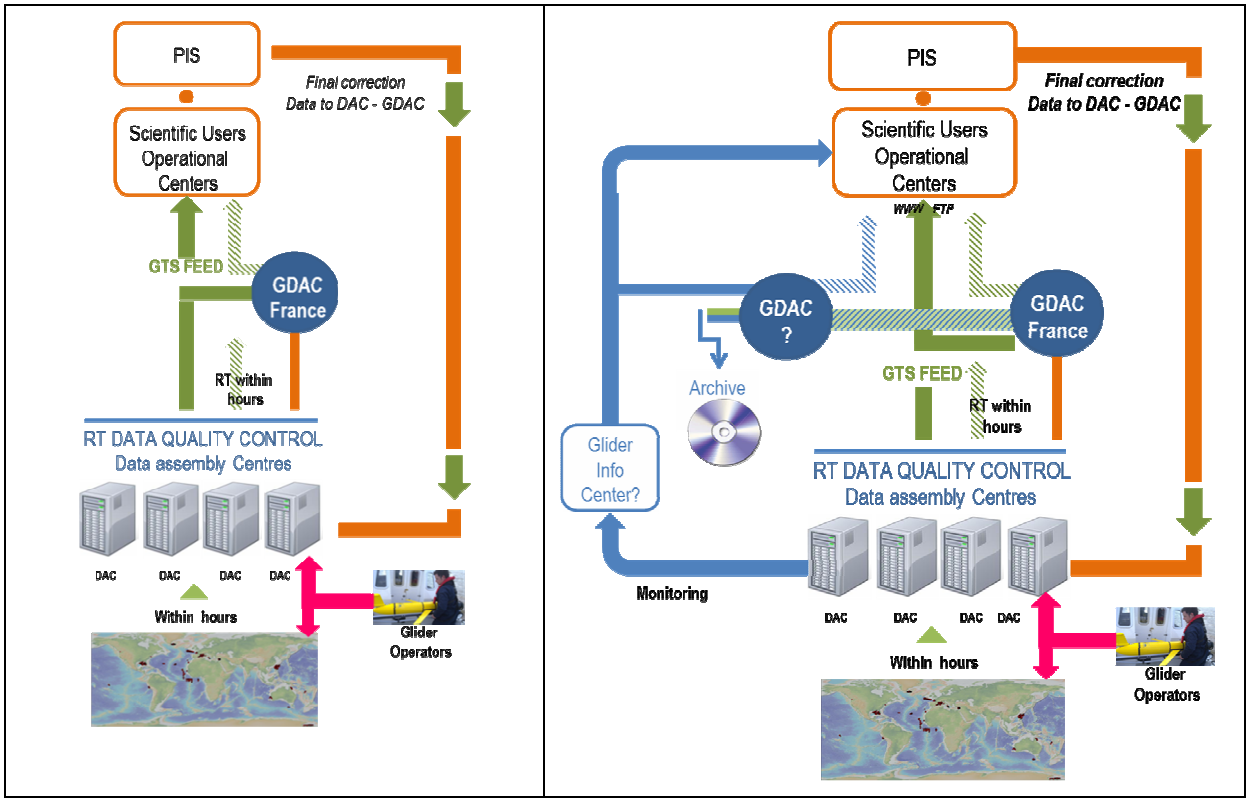
The following actors have been identified with the following duties:

- **Glider Operator:** Team in charge of steering the glider, collecting all the metadata and the deployment information required for processing, collect all the data transferred in real-time by the glider. Collect the post-recovery high resolution data.
- **PI : Principal Investigator:** Team or scientists who define the glider mission, deploy the glider and carry out post-recovery delayed mode QC that will need to be delivered to the research users within a few months of observations.
- **DAC : Data Archiving Center:** The DAC is the facility set up by one or more nations/institutes to provide Real-Time and Delayed mode glider data to the users. It
  - ⇒ collects the data from the Glider Operator,
  - ⇒ converts to standard exchange format,
  - ⇒ applies standardized real-time quality control,
  - ⇒ delivers data to the GTS and GDACs within few hours of the surfacing and to PIs on a more relaxed schedule,
  - ⇒ coordinates glider data handling for the gliders under their control.
- **GDAC : Global Data Centre .** The GDAC operates the data services where the master copies of the data reside. It doesn't perform any additional individual glider QC activities.
  - ⇒ Central point for data distribution on Internet for all GROOM gliders
  - ⇒ Can perform data format transformation, of set up additional services (OGC viewing service, OpenDap/Oceanotron download services…) to fulfil additional needs.

V. GROOM DATA MANAGEMENT SYSTEM

The data management system for Glider data has been derived from the one set up by Argo International for profiling floats. It relies on an agreed data flow between the actors listed previously, on a common data exchange format between DACs and GDAC and on a common set of RTQC procedures that will be applied by each DAC before providing the processed glider data to GDAC. There is also a target to homogenize the DMQC procedures but this is a longer term objective and will probably be finalized after the end of GROOM project.

The target data management system is summarized in the following figure on the left as well as its possible extension to International framework on the right.



The list of DAC/Glider Operators/PIs was reviewed and is summarized in the following table

Country	DAC	Glider Operator	Delayed Mode PI
UK	BODC	UEA , SAMS, NOC CEFAS, BAS	UEA , SAMS, NOC, CEFAS, BAS
France	CORIOLIS	LOCEAN, LOV, MIO, LEGOS, LPO, DT-INSU	LOCEAN, LOV,MIO, LEGOS, LPO
Italy	CMRE	CMRE	CMRE
	OGS/CORIOLIS	OGS	OGS
Germany	HZG	HZG	HZG
	CORIOLIS	GEOMAR,AWI	GEOMAR, AWI
SPAIN	CORIOLIS	PLOCAN	PLOCAN
	SOCIB	SOCIB, CSIC	SOCIB, CSIC
Norway	UiB/IMR	UiB	UiB
Cyprus	OC-UCY	OC-UCY	OC-UCY
Greece	CORIOLIS	HCMR	HCMR
Ireland	BODC	MI	MI
Poland	CORIOLIS	IOPAS	IOPAS

VI. COMMON DATA EXCHANGE FORMAT

Based on data format developed by other programs (Argo, OceanSites, SeaDataNet) as well as the format used by IMOS/Australia glider centre (ANFOG) a common data exchange format was defined for GROOM and the documentation of the first version was finalized early 2013 based on the IMOS/OceanSites like format with a number of enhancements to disseminate glider data as time series and add the necessary metadata. The following enhancements were agreed and discussed at the meeting:

- Introduce SeaDataNet compliance (parameter names, attributes)
- Ensure CF-compliance.
- Identify key additional metadata e.g. recovery and deployment cruise information.
- Include configuration and technical parameters.
- Add an extra dimension for spectra data e.g. ADCP.

To properly handle the glider data, a unique platform code within EGO (mentioned in the format as a naming authority in the data) was agreed . This unique platform code is managed by P Testor at LOCEAN for EGO. By adding the deployment code, unique for a platform, a unique Id for a deployment (period between the launch and recovery or loss of a glider) is defined.

In order to help users to easily know what type of data is available within a glider deployment file the following processing level have been defined

Value	Meaning
R	Real-time data. Data coming from the (typically remote) platform through a communication channel without physical access to the instruments, disassembly or recovery of the platform. Example: for a glider with a radio communication, this would be data obtained through the radio.
P	Provisional data. Data obtained after the instruments or the platform have been recovered or serviced. Example: for instruments on a glider, this would be data downloaded directly from the instruments after the glider has been recovered on a ship.
D	Delayed-mode data. Data published after all calibration and quality control procedures have been applied on the internally recorded or best available original data. This is the best possible version of processed data.
M	Mixed. This value is only allowed in the global attribute "data_mode" or in attributes to variables in the form "<PARAM>:DM_indicator". It indicates that the file contains data in more than one of the above states. In this case, the variable(s) <PARAM>_DM specify which data is in which data mode.

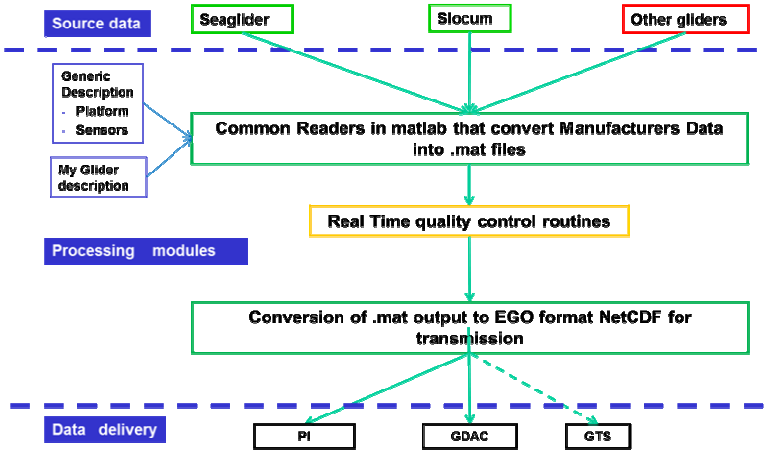
Within its life a deployment file on GDAC goes between processing levels in the following orders : R->P->D or R->M->D or R->D

The user manual describing the Glider data format is available at http://www.coriolis.eu.org/Observing-the-ocean/Observing-system-networks/EGO-gliders/Documentation

During the EU FP7 Ocean Data Interoperability Platform meetings in December 2013 Justin Buck, Thierry Carval, and Bartolomé Garau met with representative from the US Integrated Ocean Observing System (IOOS) and the Australian Integrated Marine Observing System to harmonise glider formats between continents. The meeting was successful and the revised format is expected to be published in early 2014.

VII. GLIDER PROCESSING CHAIN AT DACS

The following matlab real-time processing chain has been defined and aims at being that could be plugged after the manufacturer tools to read the seaglider and slocum files, qualify them in real-time and generate the GROOM NetCDF files.



The development of the matlab processing chain has shared among the participants to the December 2012 meeting. A collaborative development site has been set up to share softwares, documents,

configuration files on a SVN repository (https://forge.ifremer.fr/svn/oo-ego-gliders/trunk) and the issues are tracked on Mantis (https://forge.ifremer.fr/mantis).

In the proposed glider processing chain the elements dealing with reading, reformatting and RTQC are developed in a way to be shared with the GROOM/EGO community. The only part that varies with manufacturers is the data source reader that converts glider data into a .mat file. Configuration files are defined using JSON language, which is, like XML, an ASCII format but easier to manipulate for a human being and also easy to generate from configuration database. It allows to describe hierarchically a Cruise/a glider deployment / Sensor navigation and sensors/ conversion from manufacturer to EGO format. The template will be provided on the collaborative platform described earlier.

VIII. QUALITY CONTROL PROCEDURES

VIII.A. Real Time Quality Control procedures

Gliders, like other platforms such as Argo, Ferrybox, Mooring, measure ocean parameters they can transmit to shore in real-time. Among these parameters some of them are likely to be assessed in Near Real Time and used both by operational and research users. Within GROOM we decided to review the available Near Real Time Quality Control procedures that were available for Temperature, Salinity, Chlorophyll-a and Oxygen. The Near Real Time procedures will be run automatically on the data, they will allow to identify very bad data and provide this information through Quality flags positioned on each measurement. The QC Flag scale is compatible with SeaDataNet, Argo, OceanSites, GTSPP, GOSUD and MyOcean. The NRT QC procedures for T&S will be derived from Argo QC procedure and EGO QC manual taking into account the changes agreed at meeting and updates already defined with MyOcean. For Chlorophyll-A the NRT procedure will be developed jointly with BIO-Argo and are based on the development made within Pabim white book (http://www.obs-vlfr.fr/OAO/file/PABIM_white_book_version1.3.pdf). For Oxygen parameter, there was an agreement to deliver Oxygen data using a common unit DOXY in micromole/kg whatever information is sent to shore by the float. The conversion method will be the one adopted by Argo and described in http://www.argodatamgt.org/content/download/2928/21973/file/ARGO_oxygen_proposition_v1p2.pdf. The NRT procedure will also be developed jointly with BIO-Argo. When finalized the RTQC procedure will be available at http://www.coriolis.eu.org/Observing-the-ocean/Observing-system-networks/EGO-gliders/Documentation.

VIII.B. Delayed Mode Quality Control procedures

The way to handle the high resolution data, which are retrieved from the glider flash card after glider recovery, has been studied. These data need to be merged with the real time data transmitted by satellite link: the high resolution data replace the real time one. If the high resolution data are not complete (i.e. in case they were erased when the flash card got full), the real time data are kept to fill in the gaps. The first step will be to re-run the NRTQC automatic procedures on the whole time series. Then a cross calibration along the whole deployment with cruise data done at deployment and recovery will be performed and a report is issued to document the glider deployment. The following data stream was agreed: all high resolution data must be transmitted by Glider Operator to the DACs and then to PI for delayed mode QC. The high resolution data can also be transmitted to PIs in manufacturer format by the Glider Operator. The delayed mode data are transmitted back to DAC in the common GROOM format and send to GDAC by the DAC. Working group have been defined to proposed delayed mode QC procedures for the 4 core parameters identified in real time and the work is still in progress.

IX. GDAC SERVER ORGANIZATION

The GROOM GDAC (global data assembly center) is the users' access points for GROOM data. It is located in France (Coriolis, http://www.coriolis.eu.org). The GDACs handles GROOM data, metadata, and index files on ftp servers. For redundancy issues it would be good to set up another GDAC (location still to be defined). The servers at both GDACs will then be synchronized at least daily to provide the same EGO data.

The user can access the data at either GDAC's ftp site.

Since July 2013, the first glider NetCDF times-series data files from Coriolis DAC are publicly online.

For each glider deployment, a version 1.0 data and metadata file is freely available.

- http://www.ifremer.fr/co/ego/ego/v2

From these root directories of the GDACs downward, the organization of the directories and files is:

- XXX/YYY/FileName.nc
  XXX: glider code
  YYY: glider deployment code

The site codes will be listed in the "GROOM catalogue" document at GDAC's root directory.

Remark on historical data

The European gliders NetCDF data files collected before the GROOM project (Mersea, MyOcean) are still available from http://www.ifremer.fr/co/ego/ego . They will gradually be reprocessed and transferred in the above GDAC ftp server.

IX.A. File naming convention

The GROOM file names use the following naming convention for data and metadata files.

YYY/GL_XXX_YYY _ZZZ_T.nc

- GL: GROOM gliders prefix
- XXX: deployment day YYYYMMDD
- YYY: platform code from the GROOM catalogue
- ZZZ: deployment code
- T: data Mode
  - R: real-time data
  - P : provisional data
  - D: delayed mode
  - M: mixed delayed mode and real-time.
- .nc : NetCDF file suffix

Example

- PYTHEAS/GL_20100612_PYTHEAS_MooseT00_09R_R.nc

This file contains observations and metadata from the Pytheas glider, from the Moose deployment performed in June 2010.

IX.B. Index of glider deployments files

To allow for data discovery without downloading the data files themselves, an index file is created at the GDAC level, which lists all available data files and the location and time ranges of their data contents:

- The data index file is located at the root directory of the GDAC.

- The index file contains the list and a description of all data files available on the GDAC.
- There is a header section, lines of which start with # characters.
- The information sections are comma-separated values.
- Each line contains the following information:
  - file: the file name, beginning from the GDAC root directory
  - date_update: the update date of the file, YYYY-MM-DDTHH:MI:SSZ
  - start_date: first date for observations, YYYY-MM-DDTHH:MI:SSZ
  - end_date: last date for observations, YYYY-MM-DDTHH:MI:SSZ
  - southern_most_latitude, decimal degrees
  - northern_most_latitude, decimal degrees
  - western_most_longitude, decimal degrees
  - eastern_most_longitude, decimal degrees
  - geospatial_vertical_min, decibar
  - geospatial_vertical_max, decibar
  - update_interval: M monthly, D daily, Y yearly, V void
  - size: the size of the file in megabytes
  - gdac_creation_date: date of creation of the file on the GDAC, YYYY-MM-DDTHH:MI:SSZ
  - gdac_update_date: date of update of the file on the GDAC, YYYY-MM-DDTHH:MI:SSZ
  - data_mode: R, P, D, M (real-time, provisional, delayed mode, mixed; see reference table 19)
  - parameters: list of parameters (standard_name) available in the file separated with blank

The fill value is empty: "".

GDAC data files index: EGO_files_index.txt

```
# EGO FTP GLOBAL INDEX
#http://www.ifremer.fr/co/ego/ego/v2
# Contact: HTTP://WWW.EGO.ORG
# Index update date YYYY-MM-DDTHH:MI:SSZ: 2008-03-30T18:37:46Z
#
#file,date_update,start_date,end_date,
southern_most_latitude,northern_most_latitude,western_most_longitude,eastern_most_longitude,
geospatial_vertical_min,geospatial_vertical_max,update_interval,size,gdac_creation_date,gdac_update_date,
data_mode,parameters
PYTHEAS/GL_PYTHEAS_201006_R_LATEX.nc,2008-04-12T08:05:00Z,2007-03-17T18:07:00Z,2008-04-
12T08:05:00Z,0,0,-170,-170,16.7,0,550,M,2008-04-12T08:05:00Z,2008-04-
12T08:05:00Z,R,sea_water_pressure sea_water_temperature sea_water_salinity
```