US GODAE/IFREMER Data Servers
as part of the Argo data distribution network

Author : S. Pouliquen,
Date : October 2003
Version 2.4

Table of Contents

1. Overview

Following the Argo Data Management meeting held at IFREMER/Brest, the Data Management workgroup agreed to make the US GODAE Data Server and IFREMER Data Server, the two official data sources for the Argo program. I will name them the GDacs for Global Data Centers in the rest of the document. The goal is for these two servers to be fed automatically by all Dacs with the latest version of their float profiles and trajectory data and float metadata. Both servers should be updated simultaneously in order to ensure consistency between the two datasets.

Individual country agencies (some acting on behalf of more than one country) assemble the data collected from the communications system and carry out the initial processing of the data to profiles and trajectories. These agencies will be referred to as Dacs.

It was also agreed that Argo centers would use NetCDF as its distribution format for the web as well as between Dacs. Each file is the responsibility of a single Dac (i.e. the data provider) who guarantees the quality and integrity of the data. A security system is set up on each server to ensure that only the appropriate Dac is given the necessary privileges to create/modify a NetCDF file for a given profiling float.

The main objective is for the users to access a unique source of data (in this case, we have two for reliability/redundancy). A central website will provide an extensive set of tools to query, retrieve, plot and compare the profiling float data dynamically. The choice of which data server to access could be determined by its proximity to the user while attempting to alleviate the load put on either server.

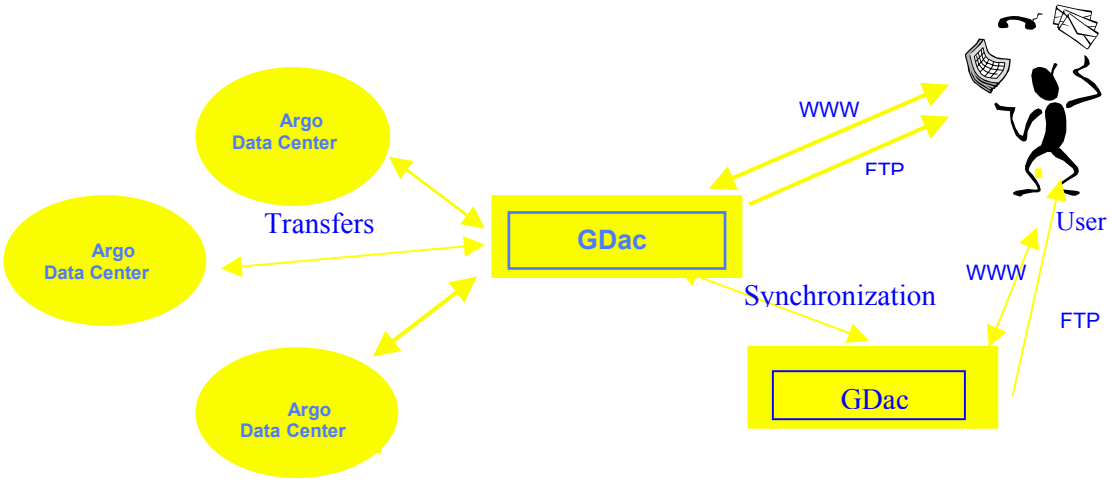
The two GDacs are expected to provide an FTP and a WWW access to the Argo data. The FTP server, suitable for data retrieval by a script/program, will provide users data organized:

- ✓ Geographically (by ocean basin) and then temporally (by year/month/day) in each basin,
- ✓ By data provider and platform (Dac and Argo float number).
- ✓ By processing date: the latest processed data organized by processing day (access to the 12 last months)

The WWW server will provide a wider set of subsetting facilities and dynamic manipulation tools; US GODAE and IFREMER agreed on a common set of WWW functionalities but may also provide additional functionalities depending of their individual choices.

2. GDac implementation

In the Argo data network, a GDac must interact with the Dacs to retrieve all the Argo data. It must then make these data available to the user community, but also to the Dacs themselves, which want to retrieve the complete Argo dataset for their studies.



The two GDacs need to synchronize periodically to be sure that they handle the same global dataset.

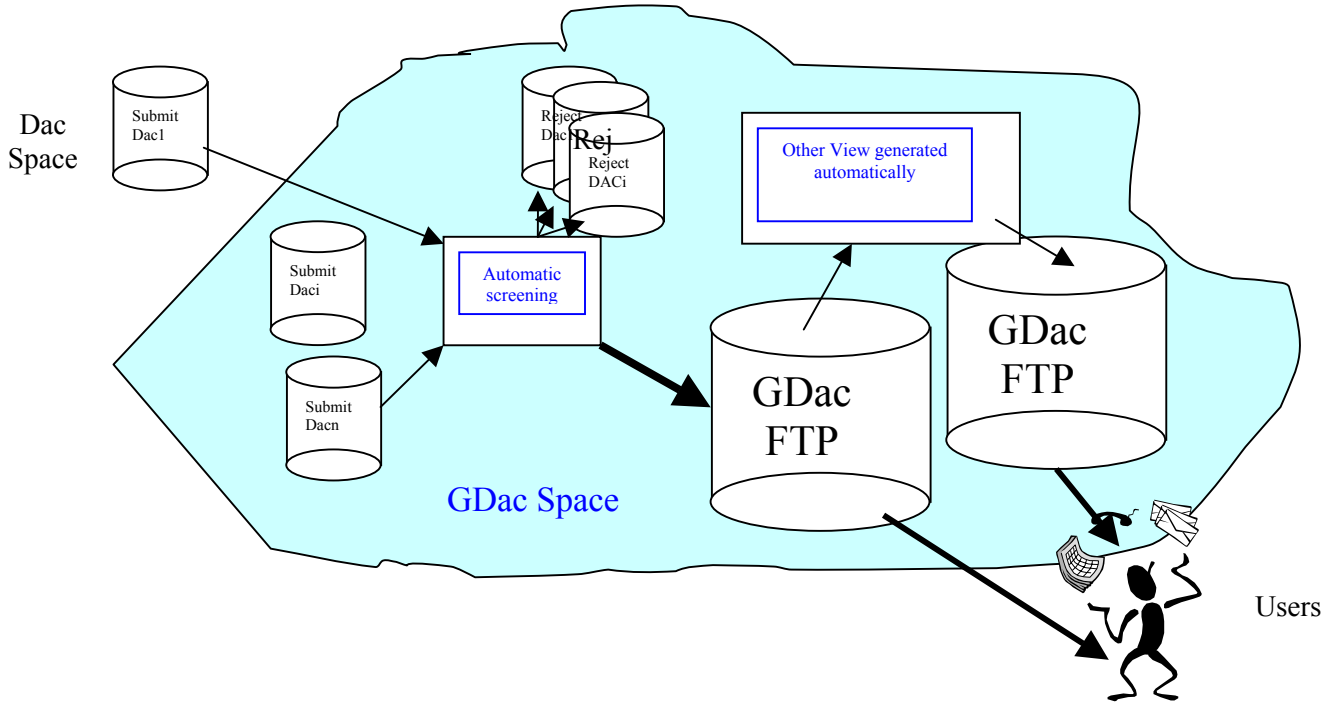
We have to consider that each GDac will elaborate value added products from this dataset; this imposes that no destruction of a float/profile should be done in the GDac space without informing the GDac manager.

The strategy is first to implement the FTP interfaces, which are crucial to collect and make available the global Argo dataset. Then for the WWW facilities, the US Godae and Coriolis data centers' managers need first to meet each other before any proposition to the Argo Data Management Workgroup.

3. Transfer between a Dac and a GDac

As we agreed on the Profile Quality Control procedures, when a file is transferred to a GDac the only controls the GDac must perform are:

- ✓ Control that the file is correct from the format point of view, in order to detect a potential transfer problem,
- ✓ Name and put the file in the adequate directory and add it to the appropriate multi-profile files in the Dac, GEO and Latest-data directories taking into account the file header information . Then, the only constraint for a Dac provider is to respect the Argo NetCDF format.



The transfers are realized through a "Submit" directory, one for each Dac, which can be located either on a Dac computer or a GDac computer. For the Dacs who have chosen to have their "Submit" directory on the GDac servers, periodically these directories are scanned and the controls described before are applied on the files . The "Submit" directory is a temporary and flat directory organization which the files only go through before being transferred on the GDac server. If the "Submit" directory is on the Dac server, the files will be transferred at the GDac before being processed the same way as described earlier.

Most of the time, the file will be correct and it will be moved to the appropriate directory on the GDac server. The organization is described in the next paragraph. This will build the "data provider and platform" view on the GDac FTP server.

A file may be rejected for two reasons:

- ✓ Because the format is not ok: an email will be sent to the Dac operator in order to retrieve a new version of this file,
- ✓ Because it's the first cycle of a new float and the metadata have not been transferred before to the GDac . In this case an email will be sent to the Dac operator asking for the metadata needed to declare this new float in the GDac database.

These files will be moved into a "Reject" directory, one per Dac, which can be examined only by the Dac operator responsible for those files. The "Reject" directories will always be on the GDac server.

The File System Organization described in the following paragraph is updated by the Dac only through the "Submit" directory (e.g. to add or change of a set of floats' profiles). To delete a profile or a float on the GDac server, the Dac operator will have to ask the GDac operator to delete it. By this constraint we are protected against an abusive deletion of profiles/or floats, which could have, because of automatic procedures, dramatic impacts on our databases and on the products generated from them.

The "Submit" directory will be used by the Dacs to replace trajectories and profiles. The GDac will use the contents of a file to determine the name to be put on a file from the "Submit" directory. This will prevent GDac from eventual errors made in a file naming which will not agree the nomenclature presented in the proposal. This will also allow GDacs to have some flexibility to improve this nomenclature with time (if necessary) without impacting the regional Dacs.

4. File System Organization at GDacs

The data provider and platform view of the GDac FTP server is organized as follows. This view is organized in separate folder for each float. Please refer to Figure1 for examples.

Profile Data

- Archived in NetCDF
- Contain the latest version of a station's vertical profiles along with metadata for that station
- Follow a naming convention =>
  - (FloatID)_prof.nc for the file containing all the profiles. This file is generated at the GDac based on data records held there
  - <R/D>(FloatID)_xxx<D>.nc for the individual profiles; xxx represents the cycle number (ex : 003). At the beginning of the name, R states for Real-Time and D for Delayed-Mode. When it's a descending profile a D is added to the cycle number.
- Follow the Argo format convention V2
- Can only be updated by the Dac responsible for the float through the "Submit" directory.

Trajectory Data

- Archived in NetCDF
- Contain the trajectory history and surface parameters collected over the lifetime of the float
- Follow a naming convention => (FloatID)_traj.nc
- Follow the Argo format convention V2
- Stored in the same folder as the station data files for that float
- Can only be updated by the Dac responsible for the float through the "Submit" directory.

Float Metadata

- Archived in ASCII
- Contain all the float-level metadata and remain fairly static over the lifetime of the float
- Follow a naming convention => (FloatID)_meta.txt
- Follow the Argo format convention V2.
- Stored in the same folder as the data files for that float
- Can only be updated by the Dac responsible for the float by email to the GDac manager.

Technical Data

- Archived in NetCDF
- Contain the technical parameters collected over the lifetime of the float
- Follow a naming convention => (FloatID)_tech.nc
- Follow the Argo format convention V2
- Stored in the same folder as the station data files for that float
- Can only be updated by the Dac responsible for the float through the "Submit" directory.

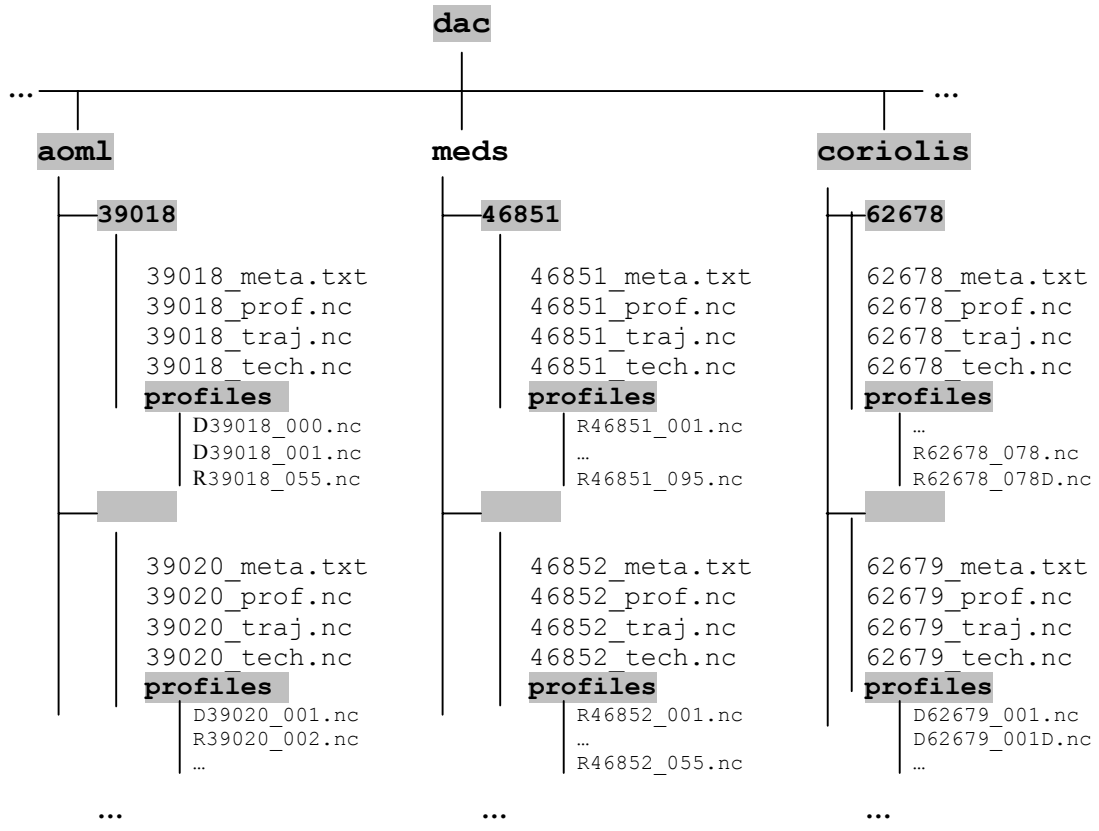


Figure 1. A sample of the hierarchical structure of files on the GDac servers.

The files are organized by data providers and float identifications as shown in **Figure 1**. This specific hierarchical structure allows Dacs to easily find a file for one of their profiling floats. PI through the

“Submit” directory a Dac will also be able to update/replace the real-time data with the better-qualified delayed-mode data.

GDacs handle two types of data :

-Real-time which are the data that come from Service Argos, are checked by the various countries and which get written to netCDF and sent to the Argo server, to PIs and to the GTS within 24 hours of the float surfacing.

-Delayed Data which are the PI checked versions: Five months later, the PIs send a QC'ed version back to the country Dacs and these data are sent on to the Argo servers.

To distinguish these two kinds of data, in the “Profiles” directories, the file name will be prefixed with R or D in the same tree. This will allow people who only want delayed mode data to get the file named "D*" through FTP and for the people who want as much data as possible to get all the data in one shot. In multi-profile files, the user will have to test the “Data_Mode” parameter contained in the files to separate Real-Time from Delayed mode profiles. In a multi-profile file, if a delayed-mode profile exists, it will replace the real-time profile.

5. GDac FTP user site for Users and Dacs

The GDacs FTP user sites are able to provide users with different views of the Argo data stored on their site.

View 1 "data provider and platform": In the previous paragraph, we have seen the "data provider and platform" view which is necessary for the Dac to be sure that they are synchronized with the GDacs at least concerning the list of profiles transferred.

View 2 "Latest received data" : Some Dacs will want to have at home a copy of the entire Argo database and so will need to retrieve regularly from the GDacs the latest data received. The GDac FTP server will provide a directory containing the latest profiles received organized by month and day of reception Figure2 . A rolling set of days will always be available on this directory (ie; one year at the beginning). This view is also useful for the users who are only interested in the latest data. There will be a file per day containing all the profiles received this day at Gdacs..

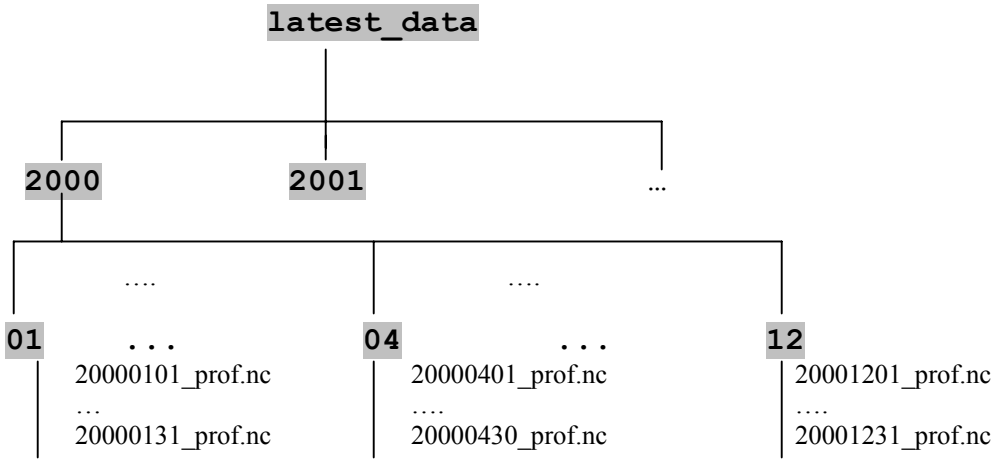


Figure 2. A sample of the "Latest data" hierarchical structure of files on the GDac servers.

View 3 "Geographical View" : The last view fulfills the needs of the users who are interested in having access to all the data of an ocean basin Figure3. The GDac provides a directory for each of the main oceans (Pacific, Atlantic, Indian). The geographical limits of Atlantic, Pacific and Indian oceans are :

- The Pacific/Atlantic frontier is 70°W.
- The Pacific/Indian frontier is 145°E.
- The Atlantic/Indian frontier is 20°E.

In each directory the data is organized by year, month and day, according to the date of the beginning of the profile. That way, people will be able to perform some selection on date.

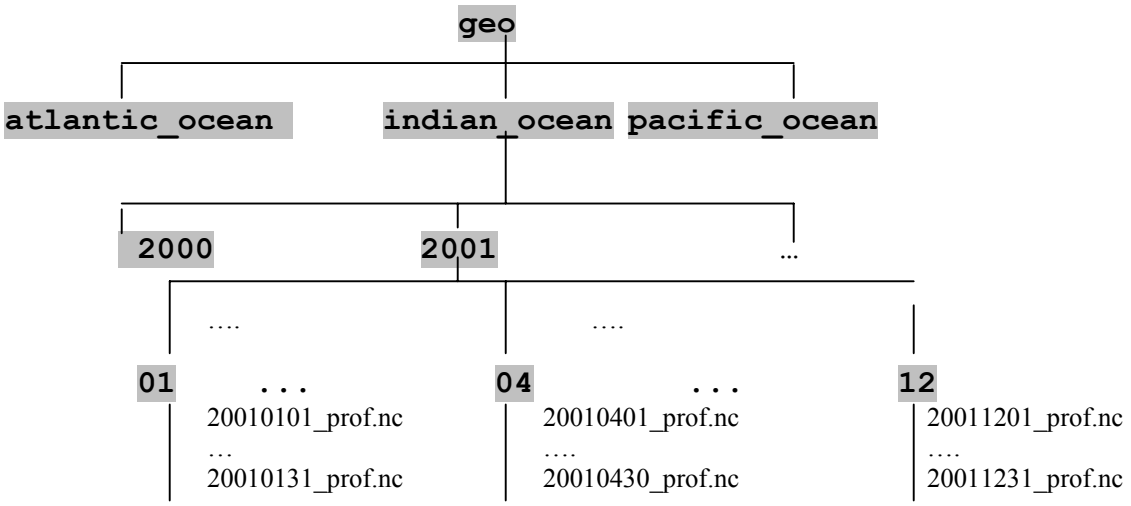


Figure 3. A sample of the "World" hierarchical structure of files on the GDac servers.

If a user wants to carry out a data retrieval based on other criteria, he will need to connect to the GDacs and structure the query themselves. Other functionalities will be provided through a WWW interface. This is presently under development

6. Index Files

At the top of the FTP sites there is 6 index files that allow users to automate data retreival from ftp site. Three of them are related to the global dataset, the three others contain the same information but restricted to the last 7 days. These files are updated each day 00UTC.

- To access metadata files: ar index global meta.xml

- To access profile files: ar_index_global_prof.xml
- To access trajectory files: ar_index_global_traj.xml
- To access metadata files of the past 7 days: ar_index_this_week_meta.xml
- To access profile files of the past 7 days: ar_index_this_week_prof.xml
- To access trajectory files of the past 7 days: ar_index_this_week_traj.xml

The complete description of these files is provided in the Argo User Manual.

7. Synchronization System

Because we will make use of two independent servers to distribute the data collected for Argo, we need a simple yet reliable method to ensure that the files located at both servers remain synchronized at all times. This of course implies that the data will not be dispatched, but rather mirrored to prevent down time as much as possible and increase the responsiveness of each server by alleviating network traffic.

All Dacs will feed both data servers simultaneously every time a new file is created or an existing file needs to be updated. This appears to be the most elegant solution as it minimizes the exchanges between data servers. In turn, it also promotes better consistency between file systems. Data should only need to be exchanged between the two Gdacs, after a period of unavailability of a GDac server or if a Dac was unable to post its files at one of the two servers.

The two GDacs servers, once a day, will compare the number and versions of their data files using the log files created by each GDac which has recorded the actions (creation, update, deletion) performed that day.

8. Actions to be set up at a Dac to upload data to Gdac

Each Dac will process and archive the data using it's own data management system. He will qualify the data according to the procedures agreed in the Argo projects and described in [QC].

Before transferring the first file, the Dac must first agree with the GDac (see Figure1):

1. either to have an account opened on the GDac servers to upload the data.
2. Or to provide an account on Dac server where the GDac can download the data.

GDac managers prefer to use the first option because it's easier for them but both options are possible. Email contacts are :

1. codac@ifremer.fr for Coriolis data center
2. Phil.Sharfstein@metnet.navy.mil for US GODAE server

All transfers between Dacs and GDacs will be made using the formats described in the [Format] documents. A format verification will be made and files which don't agree with the format will be rejected.

When a Dac wants to transfer a file 2 cases are to be considered:

1. It's the first profile of a float. In that case the Dac must first transfer the metadata file for the float and then the profile.
2. It's not the first profile, the Dac only needs to upload the latest files (profile, updated trajectory, updated technical).

The GDac will scan these directories periodically, files which have been uploaded for longer than a predefined delay (ie 1 hour at the beginning) will be checked and if the format is ok and the metadata are present at GDac, the data will be transferred in the appropriate directories on the GDac server and deleted

from the Dac account. If these conditions are not ok , these files will be moved in the Dac's "Rejected" directory and an email will be sent to the Dac contact. If a file with the same name is transferred twice, the latest file will replace the old file on the GDac server.

A Dac account at GDac is a flat directory structure where the Dac put the files plus a sub-directory "Rejected" where the rejected files are put by the GDac.

When a Dac wants to delete a file on Gdac server, there is two possibilities

1. the files are still on the Dac account and the Dac can do what he wants
2. the files are not any more on the Dac account then the Dac contact must send an email to the GDac contacts to have them deleted.

The multi-profiles files for a float will be generated by the Gdacs from the individual profiles provided by the Dac. These files will contain for each profile the original data received from the float and the best profile available. That means that a delayed-mode best profile will replace the real-time best profile when available. The original profile will always be kept in the file.

9. How to retrieve data on GDac's FTP server.

A user who wants to retrieve data from the Gdacs ftp server first have to connect to the GDac FTP servers :

- Coriolis : ftp://ftp.ifremer.fr/ifremer/argo
- US Godae : ftp://usgodae1.fnmoc.navy.mil/pub/outgoing/argo/

Depending then on what he wants to retrieve he has to go in the different directories:

- latest_data : for the latest processed data available by year/month and day
- geo: for all data organized by ocean/year/month/day
- dac: for the data organized by Dac/Float

Examples of data transfer instructions:

	Actions
Connect to a GDac server	Ftp <GDac address>
Retrieve all the profiles of one float processed by Coriolis data center	cd dac cd coriolis cd A006900045 get A00690045_prof.nc.
To retrieve all delayed-mode profiles for a float	cd coriolis cd A006900045 cd profiles mget D*.nc
To retrieve all the profiles collected in Atlantic Ocean in January 2002	cd geo cd atlantic_ocean/2002/01

	mget *
To collected all the data processed the 15th December 2001	cd latest_data/2001/12 Get 20011215_prof.nc
To retrieve all the archive together with metadata, trajectories and technical information	cd dac mget *

It case of difficulties users may contact the GDac contact point by emails.

10. Conclusions

In this paper, we describe an architecture for the management and distribution of Argo's profiling float data and metadata. This architecture is based on two data servers located at US GODAE and IFREMER whose primary goal is to deliver to the general public the most up-to-date data available within 24 hours after transmission from the float. Therefore the data would need to be uploaded to these servers at the same rate as it is put on the GTS.

We have described a simple architecture, which is able to fulfill the main functionalities, needed by the Dacs and the users. This architecture imposes no constraints on the Dacs and the GDacs internal systems. It requires read/write privileges on the GDacs systems, which should be compatible with the network security procedures of the two GDacs.

11. Applicable documents

[Format] Argo Data User Manual ; Version 2.0, October 2003

[QC] Real Time QC procedure , Version 1.0, October 2001)

[Argo-DM] Argo Data Management Handbook, Version 1.1, September 2002