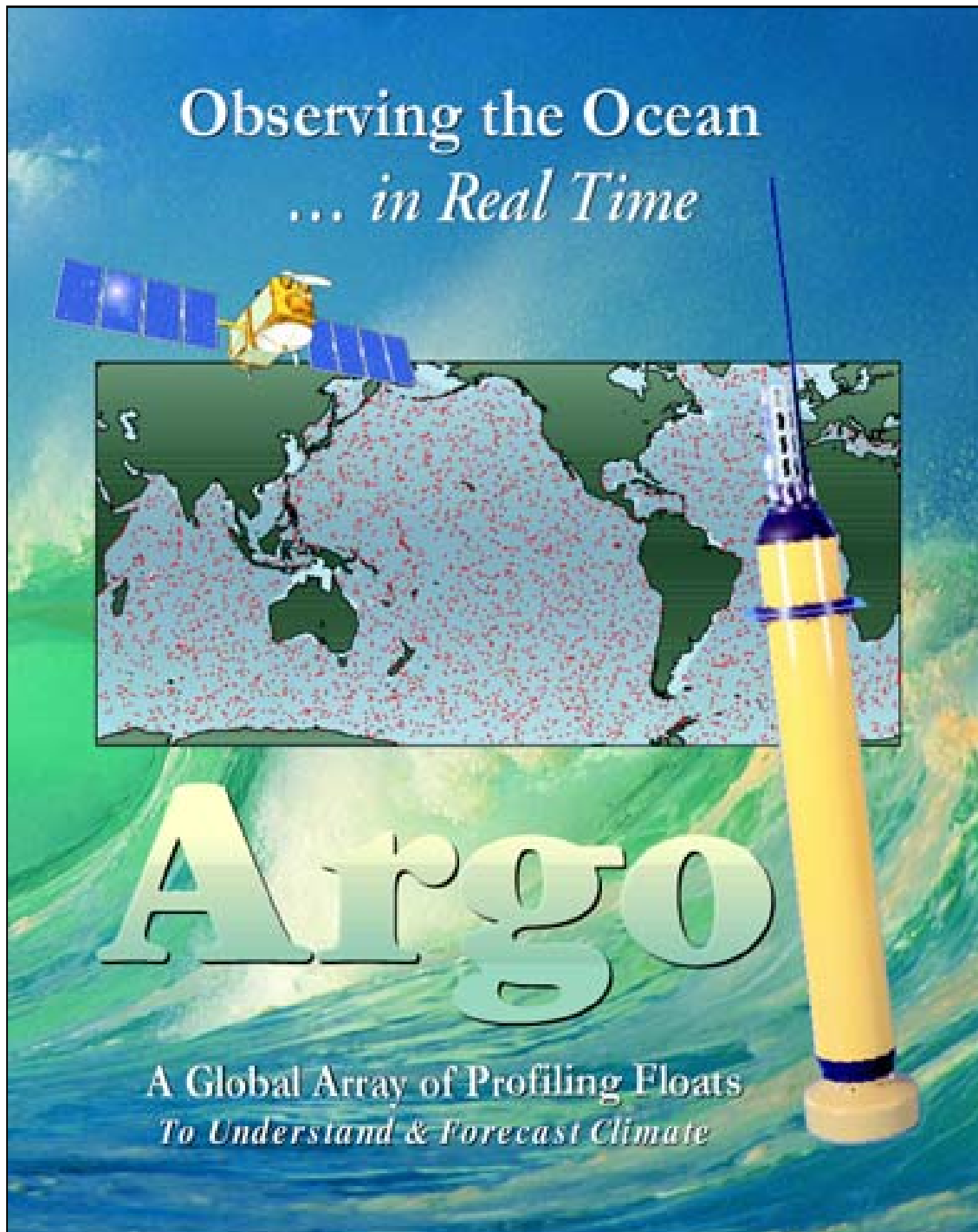


# Data Management Handbook



Last updated: July, 2002

# TABLE OF CONTENTS

<b>1. INTRODUCTION</b> .....	<b>4</b>
<b>2. GLOBAL DATA FLOW</b> .....	<b>5</b>
<b>3. RESPONSIBILITIES</b> .....	<b>6</b>
3.1. NATIONAL CENTRES: .....	6
3.2. GLOBAL ARGO CENTRES .....	7
3.3. PRINCIPAL INVESTIGATORS.....	9
3.4. REGIONAL CENTRES .....	9
3.5. THE ARGO INFORMATION CENTRE (AIC) .....	10
3.6. LONG TERM ARCHIVE .....	10
<b>4. DATA MANAGEMENT TASKS</b> .....	<b>11</b>
4.1. FLOAT IDENTIFIERS .....	11
4.2. DATA FORMATS .....	11
4.3. REAL-TIME QUALITY CONTROL.....	12
4.4. DATA TRANSMISSION.....	12
4.5. DELAYED-MODE QUALITY CONTROL .....	13
4.6. DATA VALIDATION .....	13
<b>5. GLOBAL DATA MANAGEMENT ACTIVITIES</b> .....	<b>14</b>
<b>6. REFERENCED DOCUMENTS</b> .....	<b>15</b>

## HISTORY

---

Version	Date	Comment
0.1-0.4	December 2001 to March 2002	Initial version circulating among Argo Data Management Committee
1.0	July 2002	First version taking into account the AST remarks at the 4 <sup>th</sup> IAST meeting
1.1	September 2002	AIC et R Keeley remarks

## **1. INTRODUCTION**

Argo is an internationally coordinated activity directed at characterizing both the temperature and salinity structure of the mid- and upper-ocean and the advective field at mid-depth through deployment of autonomous profiling floats. It is envisioned that the resulting temperature and salinity profiles will be used to:

- (1) initialize climate forecast models,
- (2) detect and attribute climate change effects on the ocean,
- (3) calibrate/validate satellite altimetric data, and
- (4) increase understanding of the ocean and its role in global climate.

These objectives demonstrate that Argo data are to be utilized by both operational and research communities.

The international and global nature of Argo dictates that there will be diverse float manufacturers, deployers of floats and Centres for data management. The Argo Science Team has stated that, as practical, data processing procedures should be standardized to simplify the tasks of the data users.

To maximize the effectiveness of the Argo data-set, a data management methodology is required to ensure that high quality information is readily accessible to a wide variety of users in a timely manner. This Handbook is one step in achieving this objective. Specifically, the objectives of the Data Management Handbook are:

- (1) to standardize, as practical, data management procedures among Argo Data Centres through distribution of accepted procedures and protocols,
- (2) to provide information for new Data Centres, and
- (3) to familiarize Argo data users with data management procedures and standards.

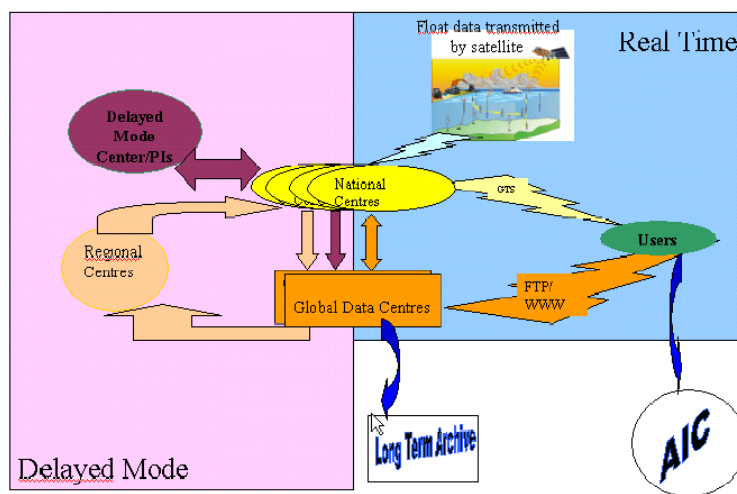
Any agency, country or consortium can take part in the Argo program. They may set up their own data processing facilities or make use of existing ones. As a contributor, they must agree that any data they collect be made available immediately and without restriction. They must also agree to abide by the procedures and practices described in this handbook.

## 2. GLOBAL DATA FLOW

The assembly of data in the Argo program is a distributed responsibility. In many cases, individual countries have established data Centres to handle the data collected by floats that their countries have contributed. In other cases, agencies within countries or groups of countries have also contributed floats to the Argo program but they make use of an existing data processing Centre.

Argo data are processed and distributed through a network involving different actors

- PI: the scientists who deploy the floats, then carry out delayed mode QC and return data to National Centres within 5 months of observations. They can delegate the task to a National Centre, but they are still the point of contact for questions about the quality of the data.
- National Centres: the data Centres who collect, qualify, process and distribute the float data for which they are responsible. Data are distributed to PIs and the GTS within 24 hours of the float surfacing. They also send the data to the Global Data Centres. National Centres are called DACs in the rest of the document
- Global Data Centres: two distribution points of Argo data distribution on Internet. They are located in Coriolis/Ifremer/France and US GODAE/FNMOC/USA. Coordination between these Centres occurs daily. Global Data Centres are called GDACs in the rest of the document.
- Regional Data Centres: Data Centres responsible for quality control on float data collected in specified regions.
- Argo Information Centre (AIC): located in Toulouse, France, responsible for providing information on the Argo program.
- Argo long term archive: data Centre located in NODC/USA in charge of ensuring the long term archive of all the Argo data.

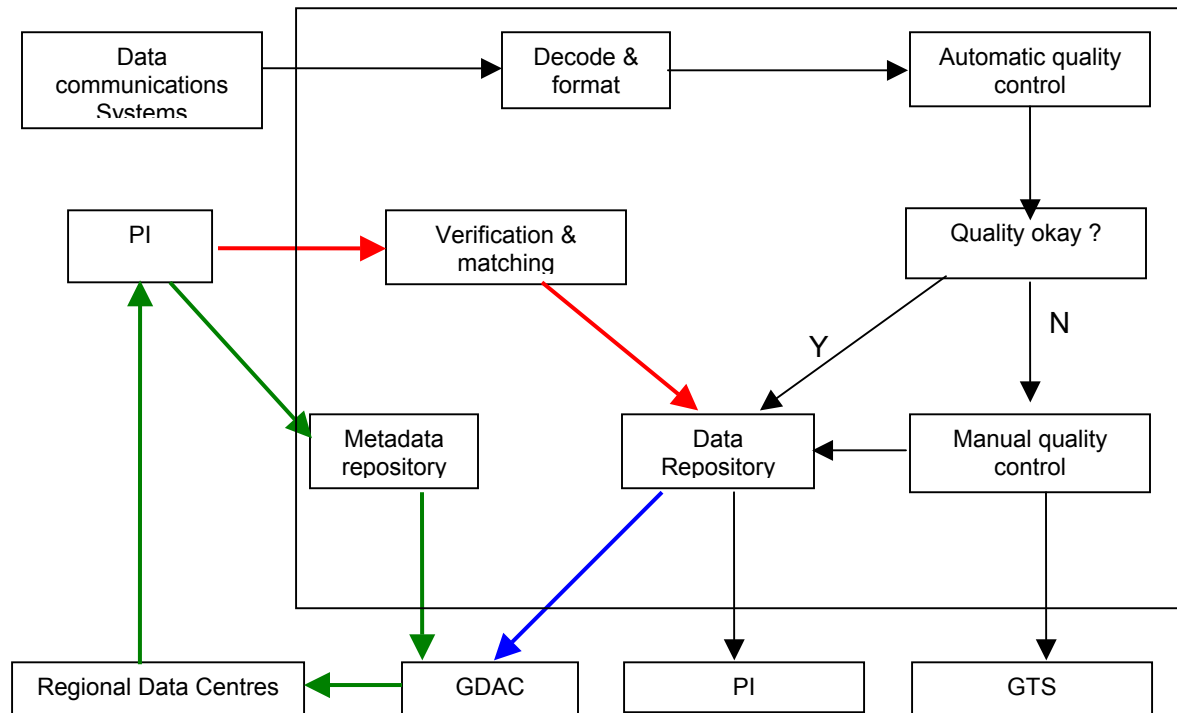


**Figure 1: A visual summary of the data flow from float to global archives**

### 3. RESPONSIBILITIES

#### 3.1. National Centres

Data processing activities at National Centres are schematized in the figure below and described in the following paragraphs.



Black line = within 24 hours

Blue Line = varying times from 24 hours to months

Green line = as needed

Red line = within 5 months

Data are transmitted from floats while they are on the surface. The satellite communications system (e.g. Service ARGOS, others) relays the data to ground stations. These either send the data to the processing National Centres, or National Centres connect directly to the communications system provider to download data. This data transfer happens several times a day.

Many of the countries supplying floats also have a National Centre designated to process the data. Although the exact nature of the work carried out at these Centres varies, each is generally responsible for converting the data stream from each float to profile and drift information. They also have the responsibility to maintain the master versions of metadata concerning the float specifications as well as data for their floats.

Data tracking is needed to assure that all components in the data system are receiving the appropriate data and at the appropriate schedules. The first step in data tracking is registration of new float information by PIs. The PI registers newly deployed floats by forwarding necessary information to the appropriate National Centre. This information is relayed to the Global Data Centres. Such information allows the National Centre to anticipate the arrival of profiles and to initiate inquiries when data are not received. Each National Centre holds the master list of instruments, profiles and current float status for the region, or PIs for which they are responsible. As part of this process, National Centres may act to request unique float numbers from the WMO before floats are deployed (see section 4.1).

National Centres prepare the data for dissemination on the GTS using a basic set of QC tests ([QC]). They also relay data to PIs within their country or who use the services of the processing Centre. At the same time, within 24 hours, the information from the float is also sent to the Global Data Centres as an alternate dissemination channel to the GTS for the research and modeling community. In some cases, the National processing Centre may also act as a distribution Centre within their own country for all Argo data collected in a region of interest to National PIs.

National Centres relay the data to the appropriate PI for delayed mode qualification. Upon receiving data from the PIs, the National Centre has the responsibility to ensure that the content of the return file is good as well as the fact that each PI made an appropriate entry (entries) in the History section of the netCDF file. If this information is missing, the Centre is responsible for working with the PI to ensure proper information is included before the data are sent on to the Global Data Centres. This transfer of delayed mode profiles should occur within 5 months of the observation date of the profiles. Additionally, if analyses at Regional Centres raise questions about the data, and if corrections are made by a PI, these pass through the National Centres back to the Global Data Centres.

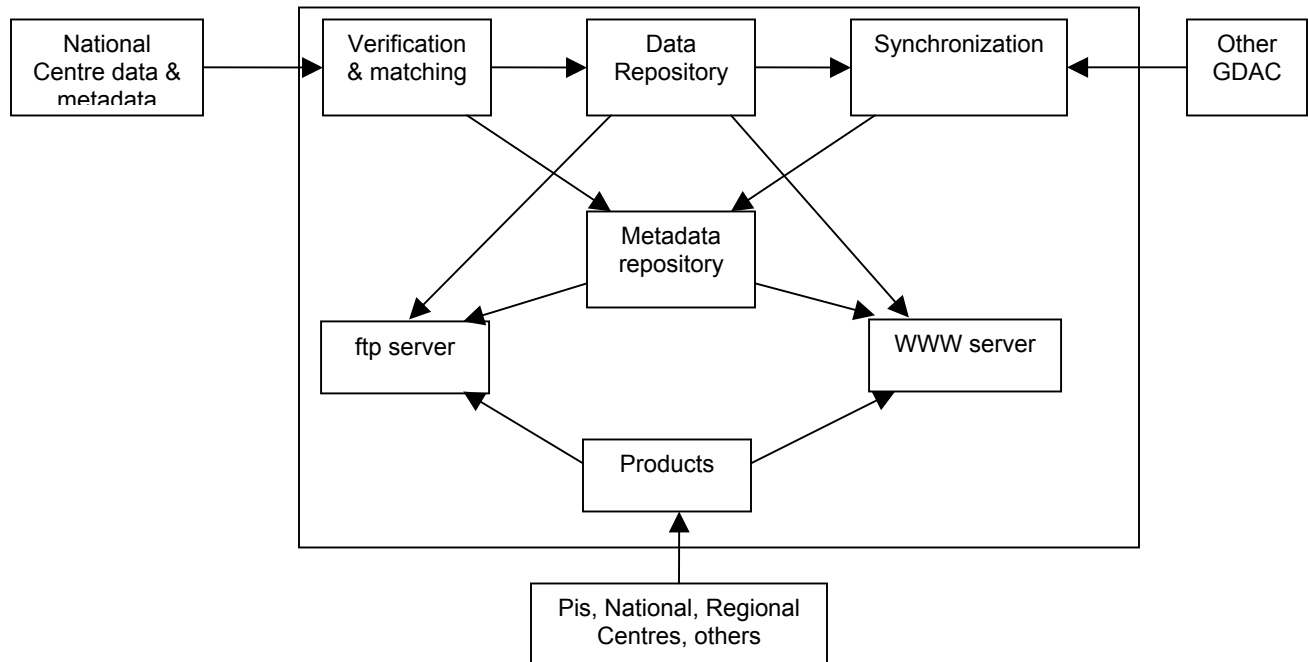
### **3.2. Global ARGO Centres**

Even though the National Centres keep the master copies of both data and metadata for the floats for which they are responsible, the Global Data Centres are the source from which all users should obtain their data. By centralizing this function, users can be assured that they are receiving the most up-to-date versions and that the data they receive is the same as what all others would receive. In addition, 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.

Data processing activities at Global Data Centre are schematized in the figure below and described in the following paragraphs.

Each of the two Global servers receives data directly from every National Centres 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. 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 GDACs are responsible for verifying the format of the files and matching received files to files already on the servers when updates are required.



The GDACs are responsible for verifying the format of the files and matching received files to files already on the servers when updates are required.

The two GDACs 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)

On the FTP site, multi-profile files will be available. These files will be generated by GDACs from the individual profile file provided by the DACs.

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 functionality depending of their individual choices.

Global Centres will also distribute to users all relevant products that may be designed or even processed by PIs or Regional Centres. This list will evolve with time and a preliminary list should be presented at the 4<sup>th</sup> Argo Science team in 2002.

### **3.3. Principal Investigators**

The PIs are the scientists who deploy the floats, who are responsible for verification of the quality of the data, and have a scientific interest in using these data.

The data sent to the PIs are the same as those sent to the Global Data Centres. PIs take the data and carry out further testing and calibrations as required to ensure high quality data. Within 5 months, they send the data back to their National Centre that then uploads the information to the Global Data Centres. A PI can delegate to another center the delayed mode QC for his floats but will keep the responsibility of guarantying the quality of his float data.

The PIs may also prepare products based on whatever instrumentation was used to collect profile data in a particular area and for a particular time frame. These may be sent to the Global Data Centres for wider distribution.

### **3.4. Regional Centres**

Regional Centres are groups that take on the responsibility, to validate all float data on a specific area (and perhaps other data as well) through more rigorous scrutiny, and to derive products. They typically retrieve data for their use from the Global Data Centres and from whatever other data servers can provide additional observations they need or want. At the 4<sup>th</sup> Argo Science Team meeting the following activities were envisioned for regional data centers:

- Determining the internal consistency of the Argo dataset by comparing Argo data from different sources in the region and through comparison with ongoing hydrographic cruises. A mechanism for feedback to PIs will be essential.
- Comparing Argo data with model output and with assimilated fields. Understanding why specific data are rejected by assimilations (model inconsistencies, systematic data errors)
- Preparing and distributing Argo data products and services.
- Providing scientific QC as a service to National programs without such capabilities.
- Coordinating Argo float deployment plans for the region. Providing advice/guidance on regional deployment needs.
- Developing new real-time quality control tests if appropriate for the particular region.
- Assembling best available recent CTD/hydrographic data for real-time and delayed mode calibration purposes.

Regional data centers may be contributed by a single National data center or may result from collaborations among two or more groups. Collaborative efforts might target different sub regions or contribute different areas of expertise. Argo National programs (and institutions) interested in forming or participating in regional data centers are listed below. The first one listed for each ocean is designated as the lead institution, to work with the others in developing a regional data center.

- Atlantic Ocean – France (IFREMER/Coriolis), U.S.A. (AOML)
- Pacific Ocean – Japan (JAMSTEC), U.S.A. (PMEL), U.S.A.(IPRC)
- Southern Ocean – U.K.(BODC), Australia (CSIRO/BOM)
- Indian Ocean – India (INCOIS), Australia (CSIRO/BOM), U.S.A.(PMEL), U.S.A. (IPRC)

In the examination of the data from a region and in the generation of these products, previously undetected problems in the data may be found. If problems are detected it is the responsibility of the Regional Centre to make appropriate notations in the files in which errors were found. Information such as the tests performed, when and by whom are important. Flags should be set on the data and if changes are made, values as received should be recorded in the History section of the data format. The data will then be returned to the PIs who will validate the corrections proposed before transmitting the data back to the Global Data Centres through their National Centre

### **3.5. The Argo Information Centre (AIC)**

The Argo Information Centre (AIC), established in Toulouse, France, is responsible for the international technical coordination of Argo, under the general supervision of the Chairman of the Argo Science Team (<http://www-Argo.ucsd.edu/>), acting in close collaboration with the Secretary of IOC and the Secretary General of WMO.

The AIC acts to resolve any issues arising between float operators, manufacturers, data telecommunication providers, data assimilation Centres, quality control and archiving agencies, WMO and IOC, encourages the participation of new countries, and assists as appropriate in the implementation of a global network.

The AIC website (<http://Argo.jcommops.org/>) gives general information on the Argo project (participating countries, contacts, real-time status of the network, status of DACs developments, maps, news, etc.) and proposes a float monitoring system which particularly permit to inform the states on the status of floats entering their Exclusive Economic Zones. The application is also used for deployment strategy.

The AIC, which doesn't distribute nor archive data, guides the users to the Argo GDACs and regional products.

### **3.6. Long Term Archive**

Argo is an internationally coordinated project that requires international participation to ensure global coverage. Individual countries have established data Centres to handle the data collected by profiling floats that their countries have contributed. The Argo program needs to establish a unique Centre that will archive long term float data in the long term. The U.S. National Oceanographic Data Centre volunteered to provide for the long-term archival of Argo data and information.

In collaboration with the Global Data Centres, it will recover all the data available and provide a long-term access to Argo data. It will also act as the Global Data Centres' backup. It will also be responsible for providing Argo data in hard copy form (such as CDs or DVDs) for users without good Internet access.

## **4. DATA MANAGEMENT TASKS**

An effective data management methodology includes many components to ensure that the highest quality data moves from sensor to user to final archive in a timely manner. It should be noted that the definition of timely is user dependent ranging from a day or less for some operational users of Argo data to months for some research users. There is also the need to provide feedback between components to allow for reporting of problems and corrections to the system. Finally, there must be a way to measure the performance of the system to ensure it meets stated requirements. Each of the following sections discusses individual components of the data system.

### ***4.1. Float Identifiers***

Each float has a platform telecommunications terminal (ptt) by which it identifies itself to the communications system. These ptt numbers are used and reused as floats are deployed and fail.

Through an agreement with the World Meteorological Organization (WMO), a unique WMO number is assigned to each float deployment. This number is of the form A9xxxxx where A represents the WMO region in which the float is deployed, 9 is a designator of a profiling float and xxxxx is a 5 digit number. For publication of data on the GTS in TESAC messages, a Q is added in front of the Id. Each time a float is deployed, a new, never before used, WMO identifier is assigned. It is this WMO identifier that is used in the data transmitted on the GTS and used as a unique identifier for a float in all data exchanges. The link between the ptt and WMO identifier is maintained in metadata files maintained at each Centre and held on the Argo Servers. When the data are transmitted on the GTS the WMO number is prefixed by the letter Q; that is, a float with WMO identifier 4900010 is identified on the GTS as Q4900010.

### ***4.2. Data Formats***

As shown in the data flow diagram, there are two ways that data can reach users. Real-time data are distributed both on the GTS and through the use of the Global Data Centres.

The data sent to the GTS are encoded, presently, in TESAC form. Data that have been flagged as wrong by the QC process are removed from the TESAC message. As well, since the TESAC code form reports depth, and also since floats report pressure, conversion from pressure to depth must be made. UNESCO software routines are recommended for this conversion (see [QC]- on QC procedures for a specific reference). Data encoded in the TESAC should be at the same vertical resolution as reported through the communications system from the float. Information about the TESAC code form can be obtained by contacting the WMO or visiting [http://www.meds-sdmm.dfo-mpo.gc.ca/meds/Prog\\_Int/J-COMM/J-COMM\\_e.htm](http://www.meds-sdmm.dfo-mpo.gc.ca/meds/Prog_Int/J-COMM/J-COMM_e.htm)

Full resolution profile data, together with quality flags, metadata, and technical data are encoded into netCDF, a self-describing format and sent to the Global Data Centres. Information about netCDF can be obtained from <http://www.unidata.ucar.edu/packages/netCDF>, the Unidata home page for netCDF. The format description of the data sent to the Global Data Centres is described in [Format]. All data reported regardless of the data quality flag assigned by the QC process are included in the netCDF file as well as the data quality flags assigned by the QC process. In addition, the netCDF file format has a section (History section) used to record which agency carried out which actions on what date and to record what QC tests were performed and which were failed at each Centre. Finally, the netCDF format stores data measured on pressure levels. It is important to be sure that the values of the depths reported in the TESAC code form are not included directly in the netCDF format as pressures.

### **4.3. Data Transmission**

Within 24 hours of data collection, the profile data are inserted onto the GTS. In future, as the GTS format is developed, it will be possible to include more metadata, including data quality flags, in the real-time data transmissions. The surface drift data and additional metadata can not initially be sent on the GTS. If for some reason, the 24-hour target for GTS insertion cannot be met, the data still should be sent to the GTS. Although some clients require the data within 24 hours, there are many others who can make use of the data after 24 hours.

At the same time as the profile data are sent to the GTS, all of the data and metadata will also go to the Global Data Centres. Because the format for storing data on the Argo servers is more capable than the initial format being used for the GTS, the Argo servers will have the most complete data and information for the floats.

Where possible, it is advisable that the data also be sent immediately to the PIs on the same time schedule. This allows them the maximum amount of time to carry out delayed mode QC.

### **4.4. Real-time Quality Control**

Because of the speed at which the data are required, the quality control procedures applied to the real-time data must run in an automated mode even though experience shows that some problems will not be caught. These problems should be caught in the delayed mode QC described in [Delayed-QC]. Each Centre carries out the same base set of tests. The data formats include a series of flags that indicate the overall assessment of quality of a data point relative to the specific tests applied.

Based on the experience acquired in the development of the Global Temperature and Salinity Profile Project, GTSPP, a set of automatic tests had been defined (see [QC]) These tests are applied to the profile data before they are converted to a message format suitable to the GTS. In the case of TESAC, since it is not possible to include data quality flags, and since it is undesirable to send data that have failed the QC tests (those that are deemed back by the automatic test procedures), any data value flagged as bad is removed from the TESAC report. If float positions, dates and times or identifiers are in question, none of the data from the float will go to the GTS.

Certain profiles may have sufficient problems that are detected by the automated testing and result in the profile not going to the GTS. In this case, it may be possible to carry out further testing, and salvage all or part of the data. If this is possible, the data should be put onto the GTS even though they fail to meet the 24-hour target described earlier. Likewise, all of the data with all appropriate flags should be sent on to the Global Data Centres for further distribution.

#### **4.5. Delayed-mode Quality Control**

Following the real-time QC, a higher level of QC is carried out by the responsible PIs and Regional Centres. The delayed-mode QC is completed by the PIs within 5 months of data collection. In delayed-mode, the responsible parties are in possession of a more complete data set for mapping and comparison purposes. Neighboring floats from other countries, different data types and subsequent profiles from a given float can all be used in the identification of suspect data.

The individual PIs role is to make a careful examination of profiles and to render her/his opinion on data integrity, salinity calibration, changes, etc.

The regional Centres are to take a larger perspective (e.g., basin-wide) for the same purpose. As with the real time QC procedures, the PI's and regional Centres attach unique QC flags and also attach recalibration data as needed to the metadata files. After QC, the data are resubmitted to the appropriate PI for verification and then to the National Centre that uploads the data to the Global Data Centres. Once agreed, the delayed-mode QC procedures will be described in [delayed QC]

#### **4.6. Data Validation**

Changes in the mix of data types in the data archives and even technology advances within a particular type can result in spurious climate signals when the data are examined. For example, it is known that XBT data contain systematic errors not present in floats and other higher quality data. It will be necessary to inter-compare the results from various profiling instrumentation (CTD, XBT, XCTD, floats, etc.) to identify any systematic differences and to ensure that spurious signals are not introduced into the long-term record. Validation reports will be available to the Argo community through the AIC. Producing such reports is one of the deliverables expected from regional Centres.

## **5. GLOBAL DATA MANAGEMENT ACTIVITIES**

To have an objective overview on Argo program progression in general, and data management activities in particular, there is the need at a global level to set up some monitoring indicators. These are not yet defined but the important elements can be highlighted now.

Continuous monitoring of the locations of operating floats is needed to ensure that design requirements are being met. Specifically, data void areas will be identified and appropriate action will be taken to fill them. While this monitoring can be carried out by individual processing Centres with access to the complete data collection, maintenance of the measurement network requires close cooperation between Argo partners.

Other types of monitoring that should be undertaken include

- testing the ability of the data system to deliver real-time data within 24 hours of the float coming to the surface,
- checking that delayed mode data are available to users within 5 months,
- checking float lifetimes,
- notifying countries of floats that may drift into their territorial waters,
- monitoring the general quality of data from the processing Centres.

In addition, the robustness of the QC procedures must be verified through inter-comparison. One approach is for the National Centres to exchange profiles, qualify the raw data and compare flags. Furthermore, the use made of the Argo data by climate models must be evaluated. Data discards must be examined to determine if problems are with the sensors or assimilation techniques.

The monitoring can be carried out by volunteer National Centres, Global Data Centres or by the AIC, and results may be made available either through the AIC and/or the Global Data Centres.

## **6. REFERENCED DOCUMENTS**

[Format]	Argo Data User Manual ; Version 1.0, July 2002
[QC]	Real Time QC procedure, Version 1.0 October 2001
[GDAC]	US GODAE/IFREMER Data Servers as part of the Argo data distribution network, Version 2.2, February 2002
[Delayed-QC]	Delayed-mode QC procedures (To write when agreed)