

Identification and citation in Euro-ARGO

Context of identification and citation in Euro-ARGO

Summary of Euro-ARGO requirements for identification and citation

Detailed requirements

The following information, based on the processing of the questionnaire by the go-between, was added by the topic leader

IDENTIFICATION

1. What granularity do your RI's data products have:

a) Content-wise (all parameters together, or separated e.g. by measurement category)?

b) Temporally (yearly, monthly, daily, or other)?

c) Spatially (by measurement station, region, country or all together)?

Euro-Argo data are available from various services with various granularities listed here: <http://www.coriolis.eu.org/Data-Products/Data-Delivery/Argo-floats-interoperability-services2>. The main access (ftp server) is <http://www.argodatamgt.org/Access-to-data/Access-via-FTP-on-GDAC>. From the ftp server, there is no geographic restriction (global ocean coverage), no temporal restriction (data continuously updated in real-time and delayed mode), no parameter restriction (all observations are available).

2. How are the data products of your RI stored - as separate "static" files, in a database system, or a combination?

Euro-Argo data are available from both "static" files or from data bases. There is one Argo NetCDF file per profile (observations in one place and time on the water column). The million profiles is also available in an aggregated way from a Thredds server. It is also available from the Ifremer-Coriolis Oracle database.

3. How does your RI treat the "versioning" of data - are older datasets simply replaced by updates, or are several versions kept accessible in parallel?

A new version of data file will update and replace the previous version. The Argo main access provides the latest version of data files. Once a month, a snapshot copy of the Argo data files is performed. A DOI (Digital Object Identifier) is assigned to the snapshot. Each snapshot archive is available online from: <http://www.argodatamgt.org/Access-to-data/Argo-DOI-Digital-Object-Identifier>

4. Is it important to your data users that

a) Every digital data object is tagged with a unique & persistent digital identifier (PID)?

b) The metadata for data files contains checksum information for the objects?

c) Metadata (including any documentation about the data object contents) is given its own persistent identifier?

d) Metadata and data objects can be linked persistently by means of PIDs?

Scientific users appreciate the DOI to cite and refer to the actual version of Argo data set they used for their publications. Checksum information: for each file of the ftp server, there is an md5 signature (checksum). When downloading data files, a user can compare his file's checksum with the server checksum on: <ftp://ftp.ifremer.fr/ifremer/argo/etc/md5/>

5. Is your RI currently using, or planning to use, a standardized system based on persistent digital identifiers (PIDs) for:

a) "Raw" sensor data?

b) Physical samples?

c) Data undergoing processing (QA/QC etc.)?

d) Finalized "publishable" data?

We use the DOI for published data.

6. Please indicate the kind of identifier system that are you using - e.g. Handle-based (EPIC or DOI), UUIDs or your own RI-specific system?

We use DOI (Data Object Identifier) from DataCite. <https://www.datacite.org/>

7. If you are using Handle-based PIDs, are these handles pointing to "landing pages"? Are these pages maintained by your RI or an external organization (like the data centre used for archiving)?

The Argo DOIs landing pages are managed by the French Sextant catalogue server. Sextant collects and makes available a comprehensive catalogue of referential data from marine environments. Sextant provides access to various geographical data via Web services using standards defined by the Open Geospatial Consortium (OGC) such as OGC-WMS (map services), OGC-CSW (catalogue services), OGC-WFS (web feature servers) and OGC-WPS (web processing services). <http://sextant.ifremer.fr/en/>

8. Are costs associated with PID allocation and maintenance (of landing pages etc.) specified in your RI's operational cost budget?

Yes, they are part of the operations costs.

CITATION

9. How does your “designated scientific community” (typical data users) primarily use your data products? As input for modelling, or for comparisons?

Argo data are crucial for understanding the ocean dynamics and changes. They are used directly by oceanographers. Argo data are assimilated in some ocean models such as Hycom (US ocean model), Mercator (French operational oceanography), Foam (UK-met office), MoonGoos (Medsea), etc... Argo data are a reference for ocean models that do not perform assimilation. Argo data are used to validate/calibrate the sea surface salinity provided by the SMOS (UE) and Aquarius (US) satellites.

10. Do your primary user community traditionally refer to datasets they use in publications:

a) By providing information about producer, year, report number if available, title or short description in the running text (e.g. under Materials and Methods)?

b) By adding information about producer, year, report number if available, title or short description in the References section?

c) By DOIs, if available, in the References section?

d) By using other information?

e) By providing the data as supplementary information, either complete or via a link

The chapter 1.2 of “Argo User's manual” insists on the user's obligations for citation in publications:

“1-2: User Obligations. A user of Argo data is expected to read and understand this manual and the documentation about the data contained in the “attributes” of the NetCDF data files, as these contain essential information about data quality and accuracy. A user should acknowledge use of Argo data in all publications and products where such data are used, preferably with the following standard sentence: “These data were collected and made freely available by the international Argo project and the national programs that contribute to it.” ”

We recommend the use of Argo DOI (Digital Object Identifier) for Argo documents and data citations. See: <http://www.argodatamgt.org/Access-to-data/Argo-DOI-Digital-Object-Identifier>

11. Is it important to your data users to be able to refer to specific subsets of the data sets in their citation? Examples:

a) Date and time intervals

b) Geographic selection

c) Specific parameters or observables

d) Other

Yes, it is important to clearly cite the actual data used in a publication. The monthly DOI is valid for the whole Argo data set. It does provide a mechanism for temporal, geographic or parameter restriction.

12. Is it important to be able to refer to many separate datasets in a collective way, e.g. having a collection of “all data” from your RI represented by one single DOI?

Yes, the Argo data set is a collection of observations performed by 15000 individual and heterogeneous floats.

13. What strategy does your RI have for collecting information about the usage of your data products?

a) Downloads/access

b) Visualization at your own data portal

c) Visualization at other data portals

d) References in scientific literature

e) References in non-scientific literature

f) Scientific “impact”

Once a year, the Argo data management team publishes a data management annual report. This report addresses items 1 and 2. As Argo data are freely available, it is not possible to have a comprehensive knowledge of item 3 (visualization of other portals). Once a year, the Argo science team publishes an annual report that addresses items 5 and 6.

14. Who receives credit when a dataset from your RI is cited?

- a) The RI itself
- b) The RI's institutional partners (all or in part, depending on the dataset contents)
- c) Experts in the RI's organization (named individuals)
- d) "Principal investigators" in charge of measurements or data processing (named individuals)
- e) Staff (scientists, research engineers etc.) performing the measurements or data processing (named individuals)

The credit from publications mainly goes to Argo program. Euro-Argo is the European contribution to Argo global program. When individual observations are used in a publication, the credit also goes to the float's Principal Investigator and its institution. The PI name and the institution are part of each Argo NetCDF file metadata.

15. What steps in tooling, automation and presentation do you consider necessary to improve take up of identification and citation facilities and to reduce the effort required for supporting those activities?

No answer given.

Formalities (who & when)

Go-between	??
RI representative	Sylvie Pouliquen et al.
Period of requirements collection	August 2015 - October 2015
Status	Information gathering completed