RDA Argo DOI final report

Introduction

This report briefly summarises the solution and dissemination of work on the application of a single DOI to the Argo data set that was supported by a RDA small grant. It covers the solution implemented by Ifremer and a summary of the associated dissemination activities.

Implemented solution

Argo data are collected and disseminated by Argo GDACs (Global Data Assembly Centers) through their FTP sites. The available data from these FTP sites are continuously changing as data are added and updated continuously.

To allow reproducibility of studies with Argo data, a snapshot of the entire data set is preserved every month. The snapshot contains all the Argo data available at the time of the snapshot creation. The one-month period between two snapshots was decided by the Scientific Committee of Argo; within a given month, changes to the whole dataset are not significant.

Initially, according to one of the suggestion of DataCite to include dynamic data (Group, Metadata Working, 2015), a DOI was set to describe the overall data set: it proposed access via FTP sites GDAC. In addition, specific DOIs were assigned to each monthly snapshot.

In March 2016, to satisfy the Argo group that did not want to inflate the number of DOIs and to get closer from the new recommendations of the Research Data Alliance (RDA) (Rauber, Asmi, van Uytvanck, & Pröll, 2015), a new single Argo DOI was published using SEANOE[1], an academic data publisher for marine sciences. This unique DOI quotes either the global data set or a specific snapshot. Each monthly snapshot is uploaded in SEANOE that assigns a URL and a key. The key 42350 for example, was assigned to the snapshot 2016-02-08.

The citation of the whole data set is performed by citing the new DOI without parameters:

Argo (2000). Argo float data and metadata from Global Data Assembly Centre (Argo GDAC). Seanoe. http://doi.org/10.17882/42182

The DOI Landing Page then presents general information detailed in Figure 1.

The citation of a specific snapshot is done by adding a fragment : the key preceded by the # character to the DOI:

Argo (2016). Argo float data and metadata from Global Data Assembly Centre (Argo GDAC) - Snapshot of Argo GDAC of February, 8th 2016. Seanoe. http://doi.org/10.17882/42182#42350

This DOI Landing Page then provides information about the specific snapshot (Figure 5).

In a technical point of view, the metadata of datasets published in SEANOE are stored in a database and data files are stored into an internally file system. The DOI is composed automatically by the

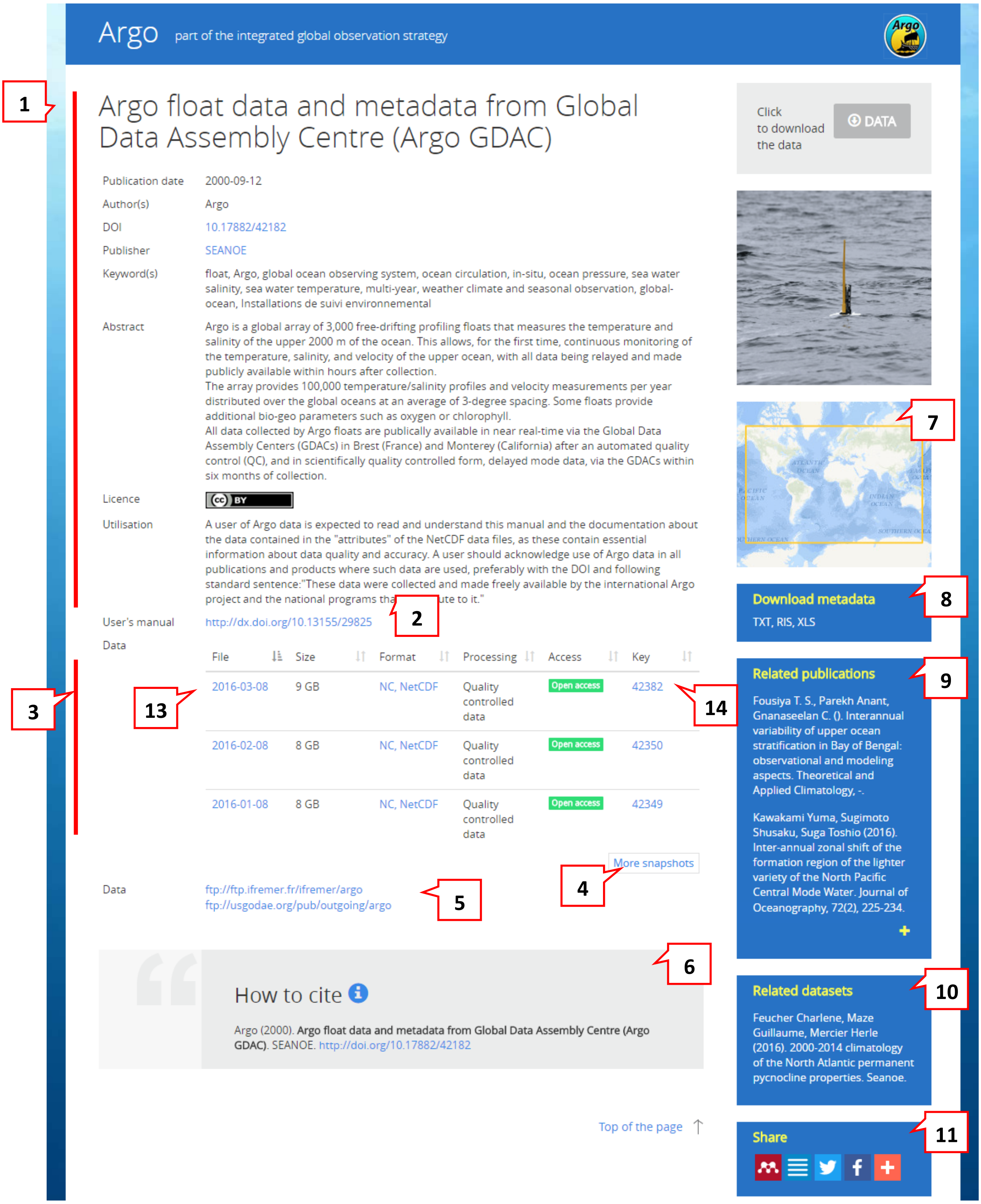
[1] http://www.seanoe.org

prefix set to SEANOE by Datacite and the automatic primary key of the dataset attributed by the database to the record (e.g. : 42182). The DOI has then so meaning: in this way, if the content of the dataset evolves in a future version, the DOI will still be appropriate. The key set to each version of the data files is the primary key of the table in the database where the list of files are saved. For the fragment too, in this way, the citation of a version has no signification (e.g. : http://doi.org/10.17882/42182#42350).

The DOI itself and even the DOI with fragments are also as short as possible: it is easier to share and to cite in a publication.

The key of the version is then associated to the DOI through a fragment (# character). This is indeed the only technical solution to propagate extra parameter trough the DOI foundation resolution service to the Landing Page. So if the http://doi.org/10.17882/42182#42350 URL is clicked, the end-user will be redirected to the http://www.seanoe.org/data/00311/42182/#42350 Landing Page. A # information can only be interpreted by the end-user browser[2]. In SEANOE, the Landing Page are static HTML pages that are updated every night with the content of the database. All the metadata specific to the different versions (e.g.: dates) are also stored in hidden fields inside the HTML. When a Landing Page is called with a fragment, the fragment will be detected by some javascript code that will update the presentation of the Landing Page automatically.

[2] To be interpreted by a server, the extra parameter should have been provided after a “?” character (e.g. : DOI?version=[version_id]) but the ? character can’t be propagated to the Landing Page by the DOI resolution service.



Argo part of the integrated global observation strategy

Argo float data and metadata from Global Data Assembly Centre (Argo GDAC)

Publication date 2000-09-12

Author(s) Argo

DOI 10.17882/42182

Publisher SEANOE

Keyword(s) float, Argo, global ocean observing system, ocean circulation, in-situ, ocean pressure, sea water salinity, sea water temperature, multi-year, weather climate and seasonal observation, global-ocean, Installations de suivi environnemental

Abstract Argo is a global array of 3,000 free-drifting profiling floats that measures the temperature and salinity of the upper 2000 m of the ocean. This allows, for the first time, continuous monitoring of the temperature, salinity, and velocity of the upper ocean, with all data being relayed and made publicly available within hours after collection.
The array provides 100,000 temperature/salinity profiles and velocity measurements per year distributed over the global oceans at an average of 3-degree spacing. Some floats provide additional bio-geo parameters such as oxygen or chlorophyll.
All data collected by Argo floats are publically available in near real-time via the Global Data Assembly Centers (GDACs) in Brest (France) and Monterey (California) after an automated quality control (QC), and in scientifically quality controlled form, delayed mode data, via the GDACs within six months of collection.

Licence (cc) BY

Utilisation 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 DOI and following standard sentence:"These data were collected and made freely available by the international Argo project and the national programs tha ute to it."

User's manual http://dx.doi.org/10.13155/29825

Data

File	Size	Format	Processing	Access	Key
2016-03-08	9 GB	NC, NetCDF	Quality controlled data	Open access	42382
2016-02-08	8 GB	NC, NetCDF	Quality controlled data	Open access	42350
2016-01-08	8 GB	NC, NetCDF	Quality controlled data	Open access	42349

More snapshots

Data ftp://ftp.ifremer.fr/ifremer/argo
ftp://usgodae.org/pub/outgoing/argo

How to cite

Argo (2000). **Argo float data and metadata from Global Data Assembly Centre (Argo GDAC)**. SEANOE. http://doi.org/10.17882/42182

Top of the page ↑

Click to download the data ⊕ DATA

Download metadata
TXT, RIS, XLS

Related publications

Fousiya T. S., Parekh Anant, Gnanaseelan C. (). Interannual variability of upper ocean stratification in Bay of Bengal: observational and modeling aspects. Theoretical and Applied Climatology, -.

Kawakami Yuma, Sugimoto Shusaku, Suga Toshio (2016). Inter-annual zonal shift of the formation region of the lighter variety of the North Pacific Central Mode Water. Journal of Oceanography, 72(2), 225-234.

+

Related datasets

Feucher Charlene, Maze Guillaume, Mercier Herle (2016). 2000-2014 climatology of the North Atlantic permanent pycnocline properties. Seanoe.

Share

Figure 1 : landing Page of the global Argo DOI

General data (Figure 1-1)

In the main part of the page, the dataset is presented with a series of general metadata (title, author, publication date, description, Creative Commons license,).

Link User's manual (Figure 1-2)

The link to Argo User's manual with the DOI assigned to that document.

List of snapshots (Figure 1-3)

By default, the last three snapshots are listed. A link (Figure 3-4) displays all available snapshots. A link to download each snapshot (Figure 3-13).

Link to FTP sites GDAC (Figure 1-5)

The link to the ftp servers of the Argo GDAC (Argo global data assemble center).

Suggested citation (Figure 1-6)

The citation proposal is built according to a format suggested by DataCite:
“Creator (PublicationYear): Title. Publisher. Identifier”

Geographical area (Figure 1-7)

The geographic coverage of the dataset.

Export metadata (Figure 1-8)

The link to download metadata of the DOI in the RIS format. This format allows automatic import in bibliographic management tools (e.g. Endnote).

List of citing publications (Figure 1-9)

The list of publications documents that cite the data DOI. The documents may also have their individual DOI. A regular monitoring is performed by SEANOE on publications which cite datasets deposited in SEANOE DOI. For Argo, this monitoring is made by the University of California San Diego.

List of associated data sets (Figure 1-10)

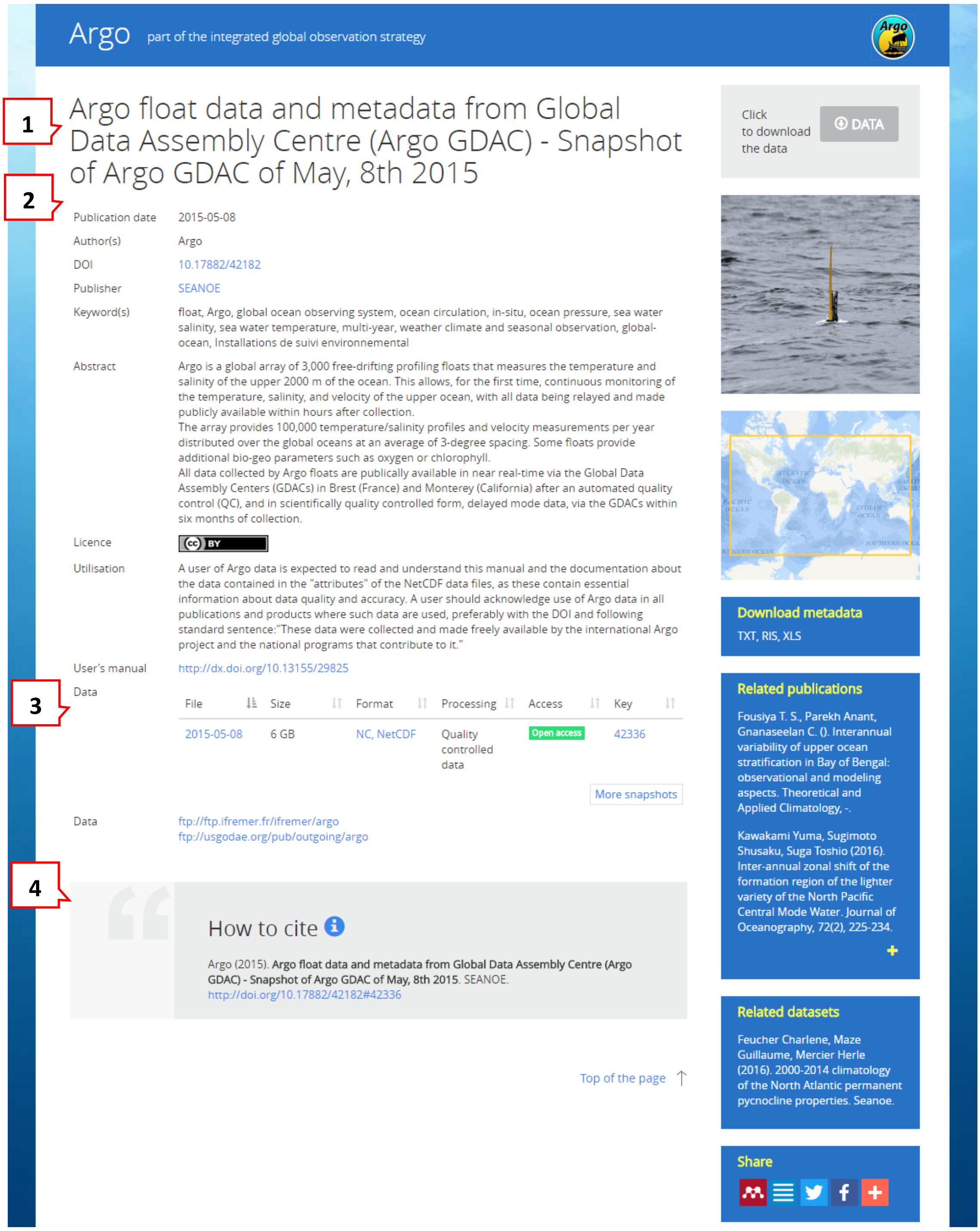
The list of related datasets that may also have their own DOI.

Link to social networks (Figure 1-11)

This link provides an automatic reporting of the dataset to social networks, through its DOI.

When the DOI Argo is called with the additional snapshot key, the following items are specifically updated on the Landing Page:

- The title (Figure 2-1)
- Date of publication (Figure 2-2)
- The requested snapshot is shown by default (Figure 2-3)
- The suggested citation (Figure 2-4)


Argo
part of the integrated global observation strategy
1
2
3
4
Click to download the data
DATA

Argo float data and metadata from Global Data Assembly Centre (Argo GDAC) - Snapshot of Argo GDAC of May, 8th 2015

Publication date	2015-05-08
Author(s)	Argo
DOI	10.17882/42182
Publisher	SEANOE
Keyword(s)	float, Argo, global ocean observing system, ocean circulation, in-situ, ocean pressure, sea water salinity, sea water temperature, multi-year, weather climate and seasonal observation, global-ocean, Installations de suivi environnemental
Abstract	Argo is a global array of 3,000 free-drifting profiling floats that measures the temperature and salinity of the upper 2000 m of the ocean. This allows, for the first time, continuous monitoring of the temperature, salinity, and velocity of the upper ocean, with all data being relayed and made publicly available within hours after collection. The array provides 100,000 temperature/salinity profiles and velocity measurements per year distributed over the global oceans at an average of 3-degree spacing. Some floats provide additional bio-geo parameters such as oxygen or chlorophyll. All data collected by Argo floats are publically available in near real-time via the Global Data Assembly Centers (GDACs) in Brest (France) and Monterey (California) after an automated quality control (QC), and in scientifically quality controlled form, delayed mode data, via the GDACs within six months of collection.
Licence	cc BY
Utilisation	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 DOI and following standard sentence:"These data were collected and made freely available by the international Argo project and the national programs that contribute to it."
User's manual	http://dx.doi.org/10.13155/29825

Data

File	Size	Format	Processing	Access	Key
2015-05-08	6 GB	NC, NetCDF	Quality controlled data	Open access	42336

More snapshots

Data ftp://ftp.ifremer.fr/ifremer/argo
ftp://usgodae.org/pub/outgoing/argo

How to cite

Argo (2015). **Argo float data and metadata from Global Data Assembly Centre (Argo GDAC) - Snapshot of Argo GDAC of May, 8th 2015**. SEANOE. http://doi.org/10.17882/42182#42336

Top of the page

Download metadata

TXT, RIS, XLS

Related publications

Fousiya T. S., Parekh Anant, Gnanaseelan C. (). Interannual variability of upper ocean stratification in Bay of Bengal: observational and modeling aspects. Theoretical and Applied Climatology, -.

Kawakami Yuma, Sugimoto Shusaku, Suga Toshio (2016). Inter-annual zonal shift of the formation region of the lighter variety of the North Pacific Central Mode Water. Journal of Oceanography, 72(2), 225-234.

Related datasets

Feucher Charlene, Maze Guillaume, Mercier Herle (2016). 2000-2014 climatology of the North Atlantic permanent pycnocline properties. Seanoe.

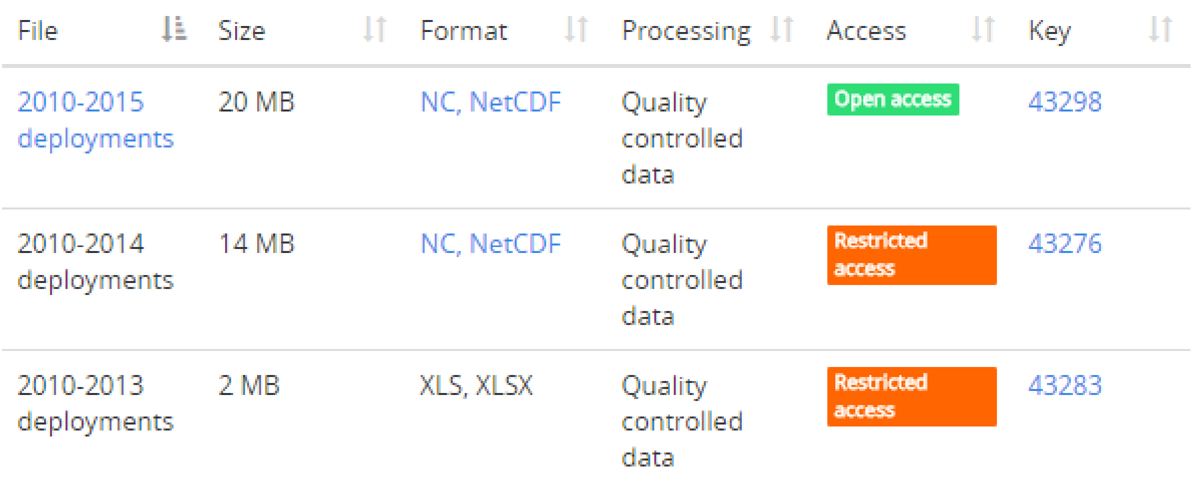
Share

*Figure 2 : Landing Page of a specific monthly Argo DOI. The #42336 identifies the monthly snapshot (http://doi.org/10.17882/42182**#42336**)*

The solution developed for publishing the Argo data in SEANOE can be implemented without additional development for any type of marine data. Since this new option is available, when authors ask how to manage versions of a dataset deposited in SEANOE, the two solutions are suggested:

- Create new DOI for each specific versions and set links between them
- Use the same DOI and offer an access to all versions in the Landing Page of the unique DOI

At the moment, all the authors have selected to manage all versions in the same DOI. Furthermore, most of them have made the choice to only offer an Open Access to the last version. The previous versions are still available on demand to ensure the reproducibility analyzes of previous articles but, in this way, the end-users are driven to the latest, most updated, data. This is the choice made for the Dyfamed dataset for example (Figure 3).



File	Size	Format	Processing	Access	Key
2010-2015 deployments	20 MB	NC, NetCDF	Quality controlled data	Open access	43298
2010-2014 deployments	14 MB	NC, NetCDF	Quality controlled data	Restricted access	43276
2010-2013 deployments	2 MB	XLS, XLSX	Quality controlled data	Restricted access	43283

Figure 3 : Data files provided by the Dyfamed dataset (http://doi.org/10.17882/43749).

Managing all versions in one unique DOI offers some advantages:

- With a unique DOI, it is easier to drive users to the latest, most updated data. If a scientist discovers the DOI of an old version in an article, will he always check that there is a link somewhere in the Landing Page to a most accurate version?
- To offer the best visibility on Google, which is one of the main source of users, it is more efficient to have only one DOI so one Landing Page and a maximum of backlinks linked to this page instead of multiple Landing Pages that share the backlinks and that provides almost the same text.
- It is easier to follow the usage and the citation of one DOI for one dataset rather than follow the usage and the citations of several DOI for one product.

However, the unique DOI strategy does not fit in all situations. For example, if the different versions have different list of authors, it may be preferable to set a specific DOI for each versions so that they can be cited differentially. Further more, for the ARGO dataset, versions are managed through snapshots of the entire dataset: this method may not be appropriate all the time especially for huge datasets because it would cost too much to save the entire dataset and not only the modifications inside the datasets.

Another solution could have consisted on using fragments to transmit a query with the DOI (http://doi.org/10.5446/12780#t=00:20,00:27). It is possible to offer the possibility to end-users to cite a DOI with parameters such as temporal extend, extraction date, geolocalization, etc, …. (e.g. http://[DOI]/#date-begin=2010-12,date-end=2015-12,date-extraction=2016-10-23). This solution would allow end-users to download and cite only a part of a dataset for example when the whole dataset is too big to be downloaded through the WEB. But this solution has its own weakness.

Sharing and citing of DOI with longs queries are less reliable: some part of the URL can be broken by articles layout. End-users can also make some mistakes in the syntax of the queries.

In 2015, the RDA have published a recommendation to cite evolving data (Rauber, Asmi, van Uytvanck, & Pröll, 2015) that offers a solution to this problem by introducing the notion of query store. The RDA especially suggests to:

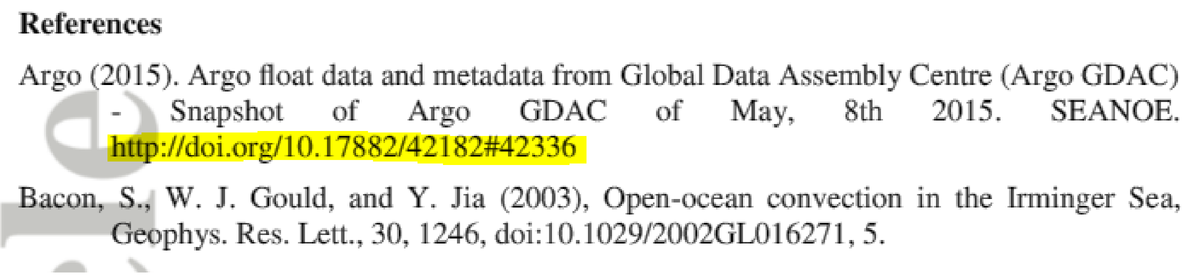
- Save all values and dates of all modifications on data (add, modification, delete)
- Save and normalize all end-users queries, save a checksum of the result, forbid modifications of the data during each query, sort results in the same way
- Set a PID (e.g. : a DOI) to each query
- Ask end-users to provide metadata citations after each query (e.g : title, …)
- Make sure to be able to re-run the queries and check the results
- Produce Landing Page with the meta-data provided by the end-users
- Delete the queries that are finally not used

The complete RDA recommendations may be complicated to implement for all dataset publishers. Furthermore, it has also its own limits too. For example, some databases are so big that exports may take a few hours. It is not possible to forbid any modifications inside the database during a few hours. When to decide that a query can be deleted because it is not used when it may takes years to exploit a dataset and publish an article about the result.

For the ARGO dataset we did not implement all RDA recommendations but the system of key set to each snapshots may be close to the RDA query store suggestion: the complexity of the export query is hidden behind a simple numeric key.

First citation of the new Argo DOI

The first citation of the new Argo DOI has been identified in Piron *et al.* (2017).



References

Argo (2015). Argo float data and metadata from Global Data Assembly Centre (Argo GDAC) - Snapshot of Argo GDAC of May, 8th 2015. SEANOE. http://doi.org/10.17882/42182#42336

Bacon, S., W. J. Gould, and Y. Jia (2003), Open-ocean convection in the Irminger Sea, Geophys. Res. Lett., 30, 1246, doi:10.1029/2002GL016271, 5.

Dissemination of results

These results have been disseminated broadly in the informatics and European research infrastructure community. The results are to be published with the paper ready for submission.

Key meetings where results have been presented are:

- RDA meeting
- Argo data management team
- Argo steering team
- IMDIS
- ODIP
- ENVRIplus meeting

The result have also been adopted within the AltantOS project for broader implementation as part of the Atlantic marine Observing System as documented in AtlantOS deliverable 7.1 (https://www.atlantos-h2020.eu/download/deliverables/7.1%20Data%20Harmonization%20Report.pdf).

References

Group, Metadata Working. (2015). *DataCite Metadata Schema for the Publication and Citation of Research Data.* doi:10.5438/0010

Piron, A., V. Thierry, H. Mercier, and G. Caniaux (2017), Gyre scale deep convection in the subpolar North-Atlantic Ocean during winter 2014–2015, Geophys. Res. Lett., 44, doi:10.1002/2016GL071895.

Rauber, A., Asmi, A., van Uytvanck, D., & Pröll, S. (2015). *Data Citation of Evolving Data Recommendations of the Working Group on Data Citation (WGDC).* Retrieved from https://www.rd-alliance.org/system/files/documents/RDA-DC-Recommendations_151020.pdf