A NEW INFORMATION AND DATA MINING TOOL FOR NORTH ATLANTIC ARGO DATA

***By G. Maze*[(1)]**

[(1)]*LPO, IFREMER, Plouzané France*

The global Argo array is made of about 3,000 free-drifting floats measuring temperature and salinity (along with possibly other parameters such as oxygen) of the upper ocean. This continuous ocean monitoring in space and time produces a tremendous amount of data, made publicly available within hours. Argo data are available for download on ftp servers hosted by the two GDACs (Global Data Assembly Centers: USGODAE and Coriolis). As of February 2012, more than 900,000 NetCDF profile files were available on these servers. From the user perspective, especially new ones, engaging with such an amount of data may be impressive, and a laborious task.

Hence, as a complementary to other online services, a new information and data mining tool for Argo data in the North Atlantic have been designed to help users manipulating the data. This is a contribution to the North Atlantic Argo Regional Center (NA-ARC) and therefore only concerns Argo profiles located in North Atlantic Ocean North of 20°S, as well as in the Mediterranean and Arctic Seas.

This new tool aims at:

- Providing an interactive user interface for Argo data mining,
- Simplifying access to information about all, or a sub-set of, profiles,
- Centralizing as much as possible information provided by other services.

Specifically, the tool is made of a website and of a web service or web API (Application Programming Interface). The website provides the user interface to services provided by the web API. The latter is public in order to allow access to services programmatically. I will now present first the database used by the system, then the website and finally the web API.

The database

Every day at 0H00 GMT, the tool scans all NetCDF profile files on the Coriolis ftp server and selects those located in the NA-ARC region. Only profiles having a POSITION_QC flag of 1, 2, 5 or 8 are selected, meaning that the position in space and time of profiles is very likely correct. A database is then created with relevant information about these profiles, such as: spatio/temporal coordinates, Data Assembly Center name, WMO (World Meteorological Organization float unique ID) and cycle number, data mode, station parameters and profiles parameters QC flags (indicating the percentage of 'good' measurements for each of these parameters). The system complements this database with additional information from other sources:

- *Quality control* information retrieved from the ftp greylist file, the LPO/Argo quality control database and the CLS/Altimetry last test results. Floats or profiles reported by one of these sources are flagged as having a "ticket" in the database. More information will be incorporated in the future, such as the Objective Analysis Warning report.
- *Descriptive* information about measurements, primarily URL pointing to figures produced by Coriolis for each parameters and profiles.

The database can be seen as a mash up of information from different sources built to provide the core information required for selection and scientific engagement.

The NA-ARC website

Without a priori knowledge, exploration of the database is made possible by the NA-ARC website available at the URL:

http://www.ifremer.fr/lpo/naarc

A snapshot of the home page is shown in figure 1 (using the Safari browser). The website provides interactive visualization tools powered by modern web technologies. A menu on the left offers six main visualization panes: "Map", "Charts", "Time Series", "Data Explorer", "Tickets and Data Quality" and a "Profiles selection wizard". A non-exhaustive list of visualization tools is: map of profiles location, pie

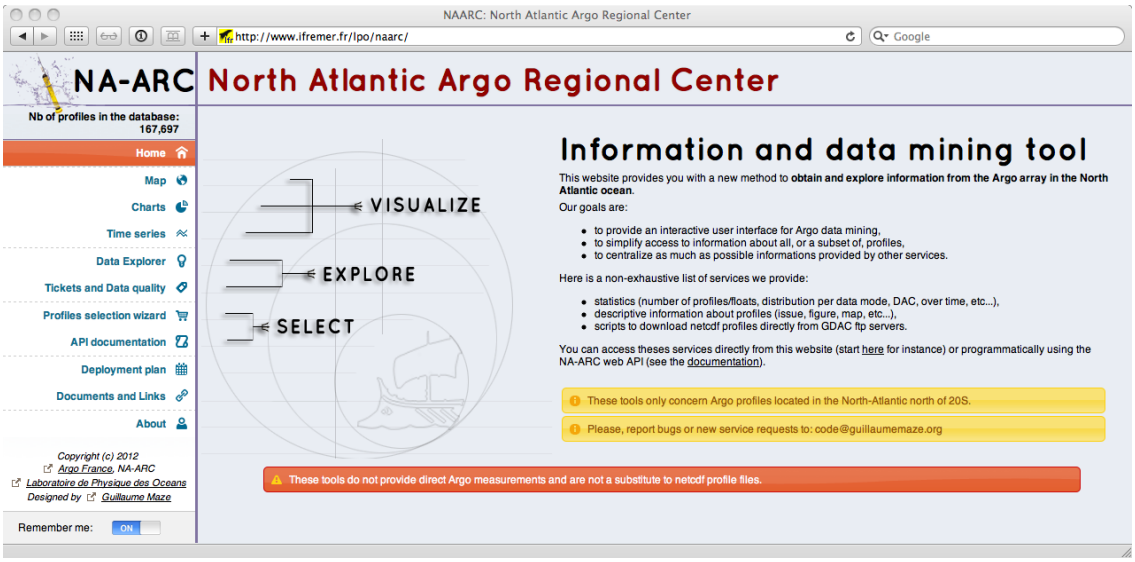

Figure 1: The NA-ARC website homepage.

charts (distribution per DAC or data mode, ⋯), time series (of the distribution per year, the seasonal cycle sampling, ⋯), gauges (for tickets, availability of oxygen measurements, ⋯), bar charts of parameter measurements quality and two search engines to retrieve tickets information and visualize figures of temperature/salinity and possibly oxygen. All plots use bright colors and scalable vector graphics so that they respond to mouse events, support animation and more importantly can be printed with optimum resolution (other elements of the website layout have also been optimized for friendly printing). Figure 2 showcases some of these tools.

Each visualization tool is embedded into an independent module developed specifically for the website. These modules provide interaction methods with the data visualized. For instance one can modify the type of chart used to represent the data, toggle between number of profiles and floats, possibly save the chart as a figure file and add restrictions to the data set used by the module to generate custom chart on demand.

By default, all visualization tools use the entire database. The "Profiles selection wizard" makes possible user specific data mining. It helps users select and define restriction parameters on profile properties in order to create a virtual sub-set of profiles (fill a "cart"). Once the sub-set is defined, all visualization tools of the website are updated automatically using the "cart" selection and temperature/salinity sections and profiles figures can be scrolled in the "Data Explorer" pane. This provides a unique way to engage with Argo data and to start exploring a sub-set of profiles without downloading any single NetCDF file. If the user is satisfied with its collection of profiles, the "Profiles selection wizard" also offers the possibility to create a script file to download NetCDF profile files from GDAC ftp servers.

One more feature offered by the website is the persistent storage of the "cart". The restriction parameters corresponding to the virtual "cart" (along with other layout parameters) are stored in the browser. This allows users to later revisit the website and eventually monitor changes to their sub-set of profiles.

The NA-ARC web API

All information provided by the website is served by the NA-ARC web API. The web API allows for users to access and mine the database from a script. This can be very powerful and provides a method for automatic processes. Here basic usage of the web API is presented. A detailed description of all available parameters and their usage, along with examples and output format descriptions can be found on the NA-ARC website online documentation.

A web API is an URL to which parameters can be added to obtain specific information from the system. The NA-ARC web API is accessible at:

http://www.ifremer.fr/lpo/naarc/api/v1/

The API parameters can be sorted in three categories:

- parameters selecting a service (get, file, qwmo, doc, plan ...)
- parameters defining the service's functions and options (n,list,coord ...),

- parameters defining restrictions on profiles properties (year,dac,box,...)

For example, querying the API can take the following form:

http://www.ifremer.fr/lpo/naarc/api/v1/?get=np

Figure 2: Samples of the NA-ARC website visualization tools.

This query calls for the function "np" (number of profiles) provided by the service "get". It returns the total number of profiles in the database (167,475 as of 2012/02/10). If one wants to obtain the number of profiles sorted by year, the option "by" can be used like: *[...]/?get=np&by=year*. This option can handle a secondary sorting key. For instance, the number of profiles per year and data mode would be retrieved using: *[...]/?get=np&by=year,dmode*. At this time, there are 9 functions available with the service 'get'. They provide an extensive list of possibilities to describe and mine the database in a simple way.

The web API default output format is JSON (JavaScript Object Notation): an easily human readable format that can be handled by scripts (python, R). It does not require any a priori knowledge to be understood. Note that the web API can also output data as Matlab evaluable strings or CSV text files.

All services and functions use the entire database of profiles by default. They can be completed by restriction parameters to select a sub-set of profiles. This is where users can fully customize their requests and express their requirements. They are more than 20 restriction parameters available at this time. They allow restrictions on *meta* information, space, time and data quality (DAC name, WMO, profile parameters, years, date range, box, tickets, parameters QC and more). As a restriction parameter example, let us consider "*around*" which allows selecting profiles near a specific one. This can be useful in quality control procedures for instance. To select profiles in a circular radius of 300km sampled ±30 days around the fifth cycle of float WMO 6900678 is as simple as adding "*around=6900678,5,300,30*" to a function call. Last, note that as many as required restriction parameters can be used to create a very specific sub-set of profiles. In the previous example, one would also select profiles with a correct quality flag on temperature using: "*around=6900678,5,300,30&temp_qc=A,B*".

Conclusion

We hope the NA-ARC website and API will provide complementary services to help users engage with Argo data in the North Atlantic. From a scientist perspective, we have tried to extend the classic engagement workflow of selection/download of profile files with a more comprehensive set of data mining and visualization tools. The entire system is flexible so that more information and services could be implemented in the future, following users suggestions.