Proceedings of the Eigth
EuroGOOS International Conference

3-5 October 2017, Bergen, Norway

OPERATIONAL OCEANOGRAPHY

Serving Sustainable Marine Development

EuroGOOS
European Global Ocean Observing System

www.eurogoos.eu

Published by:

EuroGOOS AISBL

231 Avenue Louise

1050 Brussels

Belgium

www.eurogoos.eu

To be quoted as follows:

Operational Oceanography serving Sustainable Marine Development.

Proceedings of the Eight EuroGOOS International Conference.

3-5 October 2017, Bergen, Norway.

E. Buch, V. Fernández, D. Eparkhina, P. Gorringe and G. Nolan (Eds.)

EuroGOOS. Brussels, Belgium. 2018.

D / 2018 / 14.040 / 1

ISBN 978-2-9601883-3-2. 516 pp.

TOWARDS AN INTEGRATED EU DATA SYSTEM WITHIN ATLANTOS

S. Pouliquen [1], V. Harscoat [2] and AtlantOS WP7 partners [3]

[1] IFREMER Plouzané, France. Sylvie.pouliquen@ifremer.fr.
[2] IFREMER Plouzané, France. Valerie.Harscoat@ifremer.fr
[3] AtlantOS WP7 partners. atlantos_wp7_coordination@ifremer.fr

Abstract

One goal of the H2020 AtlantOS project is to ensure that data from different and diverse in situ observing networks are readily accessible and useable to the wider community, international ocean science community and other stakeholders in this field. To achieve that, the strategy is to move towards an integrated data system within AtlantOS that harmonizes work flows, data processing and distribution across the in situ observing network systems, and integrates in situ observations in existing European and international data infrastructures or Portals, termed Integrators (e.g. Copernicus INS TAC, SeaDataNet NODCs, EMODnet, ICES, EurOBIS, GEOSS).

The actors of the integrated system, networks and integrators, are overall mature systems with long-term experience and established procedures for data collection and management. Consequently, trying to implement a sovereign and rigid set of rules for all the actors to comply with, would have been highly challenging and not in the best interest of AtlantOS. Therefore network and integrator representatives collaborated to achieve three main objectives: describe the situation at the beginning of the project, agree on a minimum set of recommendations on data harmonization and data integration, and set up a roadmap towards the integrated system.

Keywords: Data Management, Standards, Integration, EuroGOOS, JCOMM, Copernicus, SeaDataNet, EMODnet

1. Introduction

The H2020 AtlantOS project started in June 2015 and aims to optimize and enhance the Integrated Atlantic Ocean Observing Systems. One long term goal of AtlantOS is to ensure that data from different and diverse *in situ* observing networks are readily accessible and useable to the wider community, including international ocean science community and other stakeholders in this field. To achieve that, the strategy is to move towards an integrated data system within AtlantOS that harmonizes work flows, data processing and distribution across the *in situ* observing network systems, and integrates *in situ* observations in existing European and international data infrastructures and Portals, termed Integrators. The targeted integrated system shall deal with data management challenges for efficient and reliable data service to users: (1) quality control commons for heterogeneous and nearly real time data, (2) standardization of mandatory metadata for efficient data exchange and (3) interoperability of network and integrator data management systems.

1.1 The actors of the integrated system

The networks involved in Data integration for AtlantOS are Ship-based observation Networks (GO-SHIP, VOS/SOOP, CPR, fish and plankton surveys, seafloor mapping), Autonomous observing Networks (Argo, Gliders, Drifters, OceanSITES, EATN) and Coastal observing systems (Ferrybox, FOS, coastal profilers, fixed moorings).

Some Networks are organized with DACs and GDACs components. A DAC is a Data Assembly Centre typically operating at either the national or regional scale. A DAC manages data and metadata for its area with a direct link to scientists and operators, and pushes observations to the network GDAC (Global Data Assembly Centre). The GDAC is designed for a global observation network (e.g. Argo, OceanSITES, EGO for Gliders, etc) and aggregates data and metadata provided by Network DACs, in RT (Real Time) and DM (Delayed Mode).Therefore it's a portal to access, at any time, the "best version" of the Network data.

The European infrastructures or global assembly centres involved as integrators in AtlantOS are:

- For marine environmental data: SeaDataNet for validated and archived data; and the In situ Thematic Assembling Centre (INS TAC) component of Copernicus Marine Environment Monitoring Service (CMEMS) for NRT data and for the past 60 years of historical data assembled for reanalysis needs;
- For marine biodiversity data: the ICES system and EurOBIS.

The Portals involved as **integrators** in AtlantOS are:

- At European level : EMODnet lots (physics, chemistry, bathymetry, biology) fed by Copernicus Marine INS TAC, SeaDataNet and EurOBIS;
- At international level: GEOSS.

1.2 The roadmap towards the integrated system

To summarize the situation at the beginning of AtlantOS project in 2015, the data acquired by the different *in situ* observing networks were processed and distributed using different methodologies and means. Depending on the network, the data were either processed following recommendations elaborated by the network and made accessible through a unique portal (FTP or Web), or were processed by individual scientific researchers and made available through National Data Centres or at the Institution level. Some datasets were available through Integrators by ad-hoc links that were developed in past years within projects such as Copernicus, EMODNet, SeaDataNet, etc.
By exploring the data landscape, the partners in WP7 identified the needs for improvements to facilitate the access to the broad array of Atlantic observations and avoid "mixing apples and oranges", which will be to the benefit of all the actors and the users. Therefore, relying on existing European and international standards and protocols, the partners first agreed on common standards for metadata and data description (Koop-Jakobsen *et al.*, 2016), and recommendations for NRT QC of selected Essential Variables (Reverdin *et al.*, 2016). Then a data exchange backbone was designed to facilitate discovery, viewing and downloading by the users, along with tools that can be set up to at network level to plug the data on the backbone, and to facilitate integration into the Integrators. And finally it was identified that services to the users shall be enhanced to ease access to existing observations. All the guidelines for both networks and integrators are available in the Data Management Handbook (Harscoat e Pouliquen, 2016).

Time has come for the actors of the integrated system to begin the Implementation phase that aims to feed operational models and facilitate enhanced products from AtlantOS data. The networks are going to implement the recommendations for standardization across networks, plan NRT QC procedures enhancement if needed and facilitate access to network data. The integrators are going to enhance their tools for data integration (network data ingestion, viewing and downloading services on network data, cross network assessments and feedback to networks, traceability and monitoring facilities for providers and users) and to enhance the services to users (facilitate discovery through catalogues, provide OGC services, provide enhanced download facilities, facilitate visibility of existing data and provide gap identification).2. Design of the integrated system

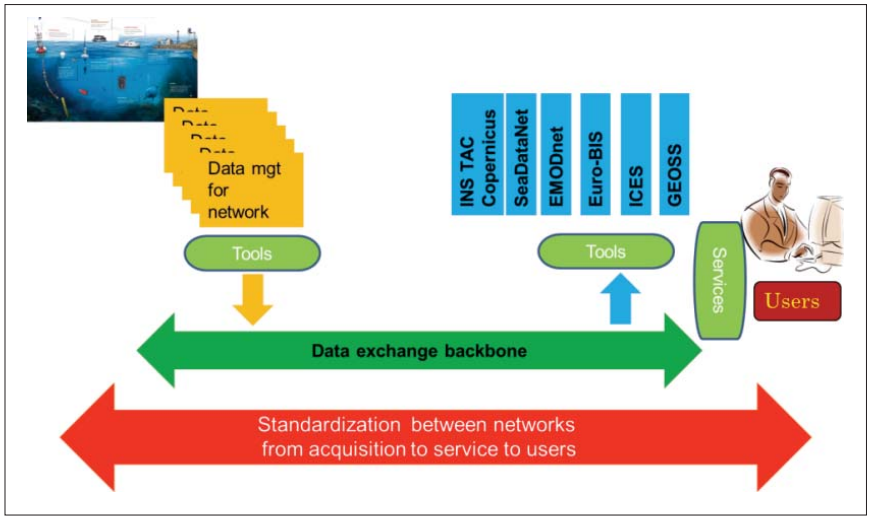

Fig. 1. The EU integrated data system, in green and red the interfaces addressed in AtlantOS.

2.1 Harmonization across networks

The actors, networks and integrators, of the future system have agreed on a set of minimum requirements and provided guidelines for best practice on essential data management aspects that will ensure cross platform coherence and facilitate better data discovery and integration. Following, the charted minimum requirements and guidelines for AtlantOS are listed.

Distribution means

As the minimum requirement for delivery service, an FTP service at the level of Network data management shall be provided. Additional services such as Web services can also be provided but are not mandatory.

Across network mandatory metadata

As a minimum, two metadata are required as mandatory in the data files and the associated platform catalogues from AtlantOS. They guarantee a continuum between data-platform-institution in an unambiguous way across the Networks and help in the traceability of the datasets. The first one is a unique ID identifying are platform and/or station identifiers and the code data provider codes. The two catalogues agreed for unique IDs management are: (1) C17 controlled vocabulary of SeaDataNet listing the codes for all platforms except stations, and (2) ICES station directory for stations. The second catalogue is more suitable as a station can be relocated and then spatial metadata are needed.

The objective of the second mandatory metadata is to give visibility to the Institutions that provide data by including in a data file the code of the Institution registered in EDMO (European Directory of Marine Organizations developed under SeaDataNet). Compared to free text, a code harmonizes the information and optimizes the discovery of datasets and allows feedback to Institutions on traceability of use.

AtlantOS set of Essential Variables

The networks in AtlantOS offer a wide variety of marine data related to many different scientific disciplines from standard parameters such as common physical ocean measurements to specialized variables such as isotopes of O_2, N_2, and Fish and Plankton surveys. An AtlantOS Essential Variables list of terms (aggregated level), related to ECV –EOV or other, has been defined and integrated in the NERC/BODC Vocabulary Server as A05 vocabulary (available online at https://www.bodc.ac.uk/data/codes_and_formats/vocabulary_search/A05/), together the mapping with existing EU (SeaDataNet vocabularies) or international (CF or WoRMS for Taxa) vocabulary standards. Each Network has to define the mapping between the metadata for the parameters in their data and the standards recommended. By doing this, a Network allows mapping on the fly without having to change its datasets.

NRT QC for selected EOVs

A core of seven AtlantOS essential variables (temperature (T), Salinity (T), Current for surface and subsurface, Sea level, Oxygen (O_2), Chlorophyll-A, Nitrate (NO_3) and Carbon (pCO_2) for surface and subsurface) are selected for implementation of common Quality Control procedures because they are acquired and controlled in NRT (24h to several days) by more than one Network among the Networks involved in AtlantOS integration activity. The recommendations have been compiled by experts on those EVs and validated by the networks acquiring those EVs and performing NRT QC. Also the harmonization recommendations include QC information to be attached to the data: both Quality flags that can be mapped to the SeaDataNet flag scale and when known processing level information ("qualified in NRT using automated procedures" or "processed in DM by Scientist").

2.2 Data exchange backbone

Platform catalogue

A simple catalogue technique was formulated (recommendation is available online: http://dx.doi.org/10.13155/45063) to be used at network global data assembling (GDAC) or portal level to (1) facilitate for users the discovery of platforms and data files, and (2) set up monitoring services. It consists of populating, continuously (creation and update) on file arrival/update, two types of indexes as simple ASCII files besides the data files made available on FTP. This technique is already implemented at the Copernicus INS TAC (content at the end of 2015: 100 000 data files and 30000 platforms). Such index files are useful for setting up synchronization between the GDAC and the user space. They can also be used to create KPIs (Key Performance

Indicators) for monitoring purposes on the networks availability, statistics on institutions or countries providing data, maps of the latest data available parameters provided delays, etc.

Detailed network and platform metadata

Concerning the metadata for platform type and sensors, it was agreed that it was an issue to be solved at Network level and that harmonization across networks was not seen as a priority. Nevertheless, a recommendation to implement SensorML for sensors whenever possible will be issued in partnership with other projects such as FIXO3, ODIP2, ENVRI+, SeaDataCloud.

AtlantOS catalogue

The AtlantOS catalogue (under construction at https://www.atlantos-h2020.eu/atlantos-catalogue) is the entry point to the integrated data system of AtlantOS. This catalogue is implemented with the GeoNetwork component of the Sextant Spatial Data Infrastructure and will also feed to the GEOSS common infrastructure.
It provides a discovery service to users and it facilitates the access to existing services (viewing, downloading and monitoring) customized to show the Atlantic Ocean as defined within AtlantOS. This catalogue firstly exposes the actors of to the integrated system and their services. It also aims to link to the monitoring services under JCOMMOPS umbrella and finally to expose the AtlantOS products. The content of the catalogue and the user services are going to be made available progressively as the integrated system is being set up, especially during the implementation phase of the project (from end of March 2017).

Implementation of data citation

To be able to operate observing systems on a long-term basis, operators are often asked to provide evidence that their platform data are essential not only for their study but also for multiple uses. Sharing data with other communities contributes to foster multiple uses of observations but makes it more difficult to trace its effective use.
In the past years, a lot of progress has been made on data citation with DOI (Digital Object Identifier). DOIs can be assigned by networks to documents and datasets for two main objectives: (1) Citation (in a publication the DOI is efficiently tracked by bibliographic surveys) and (2) Traceability (the DOI is a direct and permanent link to the document or data set used in a publication).

A guideline for the best practice for DOI assignment to AtlantOS data was formulated and is available online: http://dx.doi.org/10.13155/44515

3. Data integration and services to users

To facilitate access to AtlantOS data it has been decided to work at two levels:

- At Networks level to provide integrated access to all available data. The importance of enhancing services at the Network level is that data managers are close to platform operators and can design the system to fit the platform specificities.
- At Integrators level that build thematic services for additional targeted users and will be able to enhance their services with the help of the Networks (integration, update process, archive, etc.).

3.1 Facilitate access to Network data

The way to facilitate access to Network data for users is to set up a central point from where the data can be uploaded, or rely on existing Integrators to distribute more widely their data. This central point can be either a GDAC for the Network, or a portal with files on FTP and/or web services, allowing machine-to-machine downloading and sub-setting services.

3.2 Enhance integration in Integrators and services to users

To facilitate access to AtlantOS data it has been decided to work at two levels:

- All Integrators are planning to connect to new Network GDACs that are setting up to achieve:
  - A more complete data coverage in time and space;
  - Better quality of the integrated data as update processes will be easier;
  - Extension to more biogeochemistry data essential for Ecosystem modelling;
  - Facilitate also links between Integrators (Copernicus INS TAC <-> SeaDataNet, Copernicus INS TAC <-> EMODnet);
- All Integrators are updating their data system to implement the AtlantOS recommendations on metadata and vocabularies for parameters;
- Surveys were performed to identify the AtlantOS data that were not integrated yet, and activities are going on with Networks to improve the situation for Copernicus INS TAC, SeaDataNet and consequently EMODnet;
- Implement traceability of AtlantOS observations and use methods, and develop monitoring tools and dashboards;
- AtlantOS is contributing to GEOSS through different channels including teaming up with ODIP and in addition is promoting its use as the central hub to discover environmental data and information. However, as the GEOSS Common Infrastructure is going through a transitional phase, AtlantOS will explore the best strategy for taking new initiatives like the GEOSS European Data Hub into account.

4. Conclusion

The benefit for the Atlantic community will be at different levels. For the Network operators, it allows targeting new users through wider data availability. They also take advantages of new tools and methods to improve traceability of use and monitoring of the network data availability. Furthermore they will be able to implement internationally agreed recommendations for data citation strategy and mapping between network parameters and AtlantOS Essential Variables.

Operational users will have access to enhanced products with extended time and space coverage for present parameters (T&S Current Sea Level Wave O2 Chl) both for forecast and reanalysis, but also access to enhanced products for Ecosystem model validation. The benefit will probably be more visible in European Seas, but collaboration started also for international partnership and integration of new platforms. In addition the research community will also benefit of enhanced quality of the historical products in partnership with the Networks and the Integrators.
Finally new monitoring tools will allow the AtlantOS coordination to have more visibility on what data is freely available for users and provide inputs for the elaboration of the AtlantOS Blueprint, which aims at providing an integrated vision and plan for Atlantic Ocean observations.

Acknowledgements

As the minimum requirement for delivery service, an FTP service at the level of Network data management shall be provided. Additional services such as Web services can also be provided but are not mandatory.

References

Koop-Jakobsen, Waldmann, Huber, Harscoat, Pouliquen (2016). *Data Harmonization Report.* AtlantOS project. D7.1. https://www.atlantos-h2020.eu/download/deliverables/7.1 Data Harmonization Report.pdf

Harscoat, Pouliquen (2016), *Data Management Handbook*. AtlantOS project. D7.4. https://www.atlantos-h2020.eu/download/7.4-Data-Management-Handbook.pdf
Reverdin, Thierry, Utiz, d'Ortenzio, Bradshaw, Pfeil (2016). QC Report. AtlantOS project. D7.2. https://www.atlantos-h2020.eu/download/7.2-QC-Report.pdf