iFADO Best Practice: Recommendations on operational oceanographic data sharing.

Background

The effective implementation of a transnational and interdisciplinar data services depends on the capability of multiple systems to work with the same sets of data and metadata in an interoperable fashion. Metadata records associated with one resource should be accessible, accurately interpreted and subsequently used by systems other than from the originating organisation.

The iFADO project is dealing with data sources from different scientific disciplines including numerical modelling, remote sensing, traditional monitoring campaings and data obtained by novel technologies (i.e. gliders, minboats,etc.). In order to represent the different type of data on a single platform a set of minimum requirements should be set. The way to achieve this task is through defining standards and try to make them converge.

To achieve this objective is it needed to homogenize and develop the interconnection with the existing networks (CMEMS, EMODnet, SeaDataNet); adopting standard conventions (OGC, ISO); fulfilling the INSPIRE Directive;and increasing the impact of individual data by following FAIR data principles (Findable, Accessible, Interoperable, and Reusable).

The challenge for iFADO is to connect with different sources of information without building Ad Hoc protocols for each data source but through the establishment of standard procedures. The iFADO strategy involves:

- To encourage the use of common formats and quality standards by the local operatives thus facilitating the use of data and easing uptake into other data systems.
- To facilitate the exchange of data as well as the intercomparison among observatories.
- To benefit of existing European initiatives to access and exploit the data.

In summary, the main goal of data sharing in iFADO is to promote the standards for distribution and try to rely on European/international initiatives to host the data produced during the project.

Incorporating the data into European Initiatives require previous data formatting, quality standards, filename and variable names standards as well as metadata to describe the contained data. It also clearly identifies who/which institution is the data provider/owner/aggregator and will ensure that data is accessible once the project is finished. For these reasons, the iFADO project mission has been to spread the use international standards through the whole cycle from data acquisition to stakeholder use and ultimately to incorporation in European initiatives.

In situ data, remote sensing and numerical models represent common properties (e.g., salinity, temperature, chlorophyll a, etc.), but although they use similar data conventions such as NetCDF CF, the definition of their variables may not be fully compatible.

The incorporation of these data in the iFADO platform was done following the next principles:

- Each data generator should choose the service where the data is hosted given preference to international recognised data aggregators;
- If the data generator prefers to host the data, international standards and methodologies should be followed in order to meet the FAIR principles
- The iFADO platform should not host data from any data provider, unless necessary, but it should dynamically link with the service provider.
- Only if common file format is not consolidated or the data provider is not ready to produce international standards data then conversion and integration tools will be developed.

These principles make responsible of the data to the data generator and allow a longer and more sustainable maintenance of the data. Data and platform become independent from each other, and the disappearance of the platform does not compromise the maintenance of the data. It also allows that data can be included in several data platform.

Ocean data management

Following the recommendations included in the AtlantOS Data Management Handbook (Harscoat & Pouliquen, 2016), and the conclusions of the 2010 annual EuroGOOS meeting; all data should be supplied with good quality metadata and when combined with other data sources to derive new products further metadata will be needed.

Following the proposed structure of the EuroGOOS data management system there are:

- **Observation data providers:**

  Operational agencies, marine research centres, universities, national oceanographic data centres and satellite data centres.

- **Integrators:**

  The integrators are European initiatives with the goal to offer a one-point access for ocean data. They support both data providers willing to share their observation data, and users who want to access oceanographic data from a range of providers encompassing multiple types of data from multiple regions. Specifically, these are:

  - SeaDataNet for validated and archived data, commonly used for traditional monitoring data from cruises and monitoring campaigns.
  - Copernicus Marine Environment Monitoring Service (CMEMS) for satellite and model data. Also, as a component of CMEMS, In-Situ Thematic Assembling Centre (INS TAC) for NRT (near real-time) data and for the past 60 years of historical data assembled for reanalysis needs. This service only distributed selected products from selected service provider.
  - EMODnet (physics, chemistry, bathymetry) fed by Copernicus INS TAC, SeaDataNet. Although EMODnet has also designated data providers, it also provides a data ingestion service to support standardisation and distribution for other kinds of data.

Roadmap

Since all the data produced by the partnership cannot be distributed through the international initiatives. The project is also promoting a low-level standardization of chosen datasets delivering the end-users needs. Adoption of a low level of interoperability is achieved when a tool can address datasets using standards but there are discrepancies about internal structures of datasets, files and folders.

This implies an effort by each partner to transform its datasets adopting international standards as THREDDS, NetCDF and CF conventions for numerical models, remote sensing and novel technologies data or OGC-INSPIRE standards to visualize, download or catalogue the data set. A good description for cruise data can be found on iFADO Best Practice Ocean Underway Data Management.

Below a description of the lessons learned and good practices for numerical model, remote sensing and novel technologies are described.

Numerical Model & Remote sensing

A common practice on operational numerical modelling and remote sensing is to distribute files using NetCDF standard format through a THREDDS Data Server (TDS; Figure 1). A TDS is a web server that provides metadata and data access for scientific datasets, using OPeNDAP, OGC WMS and WCS, HTTP, and other remote data access protocols. The TDS is developed and supported by Unidata, a division of the University Corporation for Atmospheric Research (UCAR) and is sponsored by the US National Science Foundation (NSF). Among the services and technologies included in the TDS we can highlight:

- THREDDS Dataset Inventory Catalogue are used to provide virtual directories of available data and their associated metadata. These catalogues can be generated dynamically or statically.
- TDS can use the NetCDF Markup Language (NcML) to modify and create virtual aggregations of CDM datasets.
- An integrated server provides OPeNDAP access to any CDM dataset. OPeNDAP is a widely used, subsetting data access method extending the HTTP protocol.
- An integrated server provides bulk file access through the HTTP protocol.
- An integrated server provides data access through the OpenGIS Consortium (OGC) Web Coverage Service (WCS) protocol, for any "gridded" dataset whose coordinate system information is complete.
- An integrated server provides data access through the OpenGIS Consortium (OGC) Web Map Service (WMS) protocol, for any "gridded" dataset whose coordinate system information is complete. THREDDS integrates a version of NcWMS and its companion "Godiva" web application for WMS access and online visualization, developed at the University of Reading.

Catalog http://thredds.maretec.org/thredds/catalog/MOHID_WATER/PORTUGAL_0.06DEG_50L_3H/FORECAST/catalog.html

Dataset	Size	Last Modified
FORECAST		--
2018010100.nc	570.7 Mbytes	2019-11-24T15:21:15Z
2018010200.nc	570.7 Mbytes	2019-11-24T15:23:47Z
2018010300.nc	570.7 Mbytes	2019-11-24T15:26:23Z
2018010400.nc	570.7 Mbytes	2019-11-24T15:28:59Z
2018010500.nc	570.7 Mbytes	2019-11-24T15:31:33Z
2018010600.nc	570.7 Mbytes	2019-11-24T15:34:07Z
2018010700.nc	570.7 Mbytes	2019-11-24T15:36:45Z
2018010800.nc	570.7 Mbytes	2019-11-24T15:39:14Z
2018010900.nc	570.7 Mbytes	2019-11-24T15:41:42Z
2018011000.nc	570.7 Mbytes	2019-11-24T15:44:14Z

Figure 1. THREDDS catalogue for the Instituto Superior Técnico PCOMS Operational Modelling System showing the daily files in NetCDF format, file size and date of creation.

The capacity to generate scripts for automatic data download allow to develop many services and applications linked to a TDS including data viewing, processing and distribution. In the case of numerical modelling allow to generate more refined models for downscaling. In addition, WMS is a common service used to display numerical modelling results and remote sensing products on data exploring platforms such as the one used in the iFADO project (Figure 2).

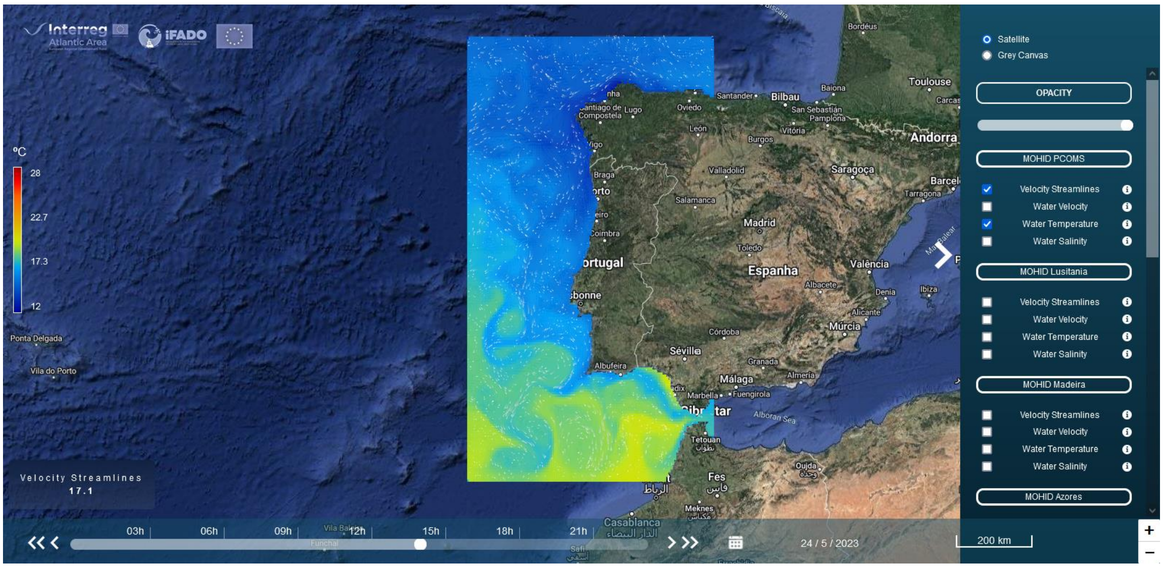

Figure 2. Capture of the iFADO web platform with the defined default layers.

Nevertheless, and despite the common use of the CF conventions, most of the data providers have a great diversity on folders organisation and naming and file and variable names. This variety hinders the capacity to find and use the information offered by these TDSs since scripts need to target specific variable names and metadata. As an example, Table 1 shows the variable names and the CF standard name given to the same property (Sea surface temperature) by different operational numerical model providers. The name is not related to the numerical engine they use, as different names can be applied while using the same numerical model.

Table 1. Sea water temperature as defined in their respective TDS by operational numerical model providers.

Institution	Numerical Model	Variable name	CF Standard Name
MeteoGalicia	ROMS	temp	potential_temperature
MeteoGalicia	MOHID	temp	sea_water_temperature
CMEMS	NEMO	thetao	sea_water_potential_temperature
MARETEC-IST	MOHID	temperature	sea_water_temperature
Marine Institute	ROMS	sea_surface_temperature	sea_surface_temperature

Once the operational service is online and distributed through a TDS, the provider has strong difficulties on adapting or modifying the name originally given. The variable names cannot be changed because operational services that are linked to these TDS may be disrupted.

A good solution has been recently proposed by INTECMAR in the context of the Atlantic Area [MyCoast project](). In that project, it has been promoted the use of a NetCDF Markup Language (NcML) (Unidata, 2023). NcML is a set of XML tags that are interpreted by the THREDDS server and some NetCDF utilities for create / modify NetCDF files. This procedure allows that the TDS count with a vocabulary that will provide to the request the original file but adding or modify some items, such as the names of their variables and their metadata. The most obvious advantage is that it is not necessary to modify the existing catalogues since NcML files are XML files pointing to the actual NetCDF file.

Autonomous vehicles monitoring

Due to the large diversity of autonomous vehicles for environmental monitoring including surface and subsurface with diverse motion mechanism data sharing protocols are not fully stablished. However, some good practices are emerging such as stablishing data aggregators and standard formats.

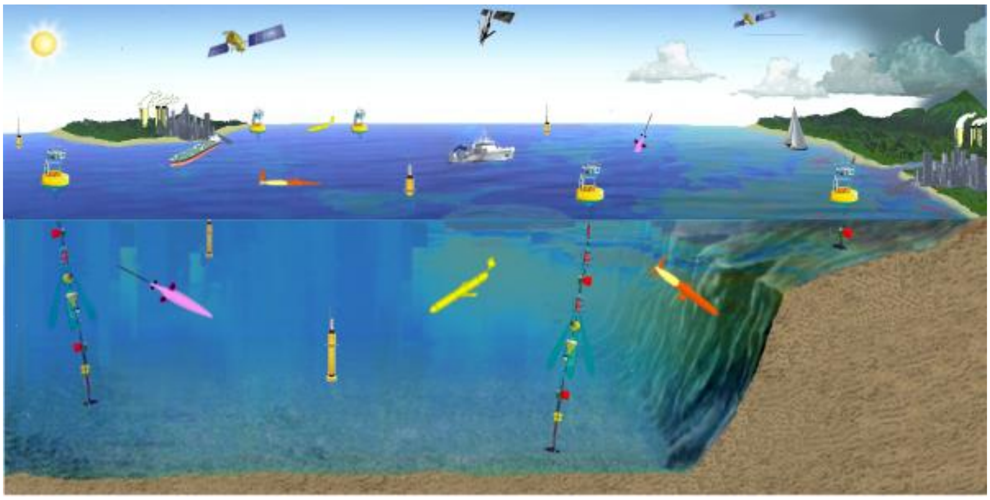

Figure 3. Marine autonomous vehicles include a large variety of surface and subsurface design with a wide range of motion mechanism (Source: EU GROOM-RI).

Probably one of the most consolidated marine autonomous monitoring systems is the Argo programme. Argo is an international programme, operational since the early 2000s, that collects information from inside the ocean using free drifting profiling floats. The floats observe profiles of temperature, salinity, currents, and, recently, bio-optical properties in the Earth's oceans. Since the data was made freely available, it set up some of the standards of data distribution.

Some autonomous vehicles such as gliders are following Argo-like NetCDF format in order to be easily managed by end-users. Common procedures for data distribution can be found in the EGO-network for gliders (https://www.ego-network.org/). During the iFADO project, partners involved in the novel technologies work package made efforts to distribute the data and monitoring location following the standards and data distribution pipeline. Figure 4 and Figure 5 represent different legs of the iFADO PAAnoramic mission operated by different institutions, Marine Institute (Ireland) and National Oceanographic Centre (UK) and distributing data with common formats and standards. This practice allowed end-user to rapidly access and use the collected data.

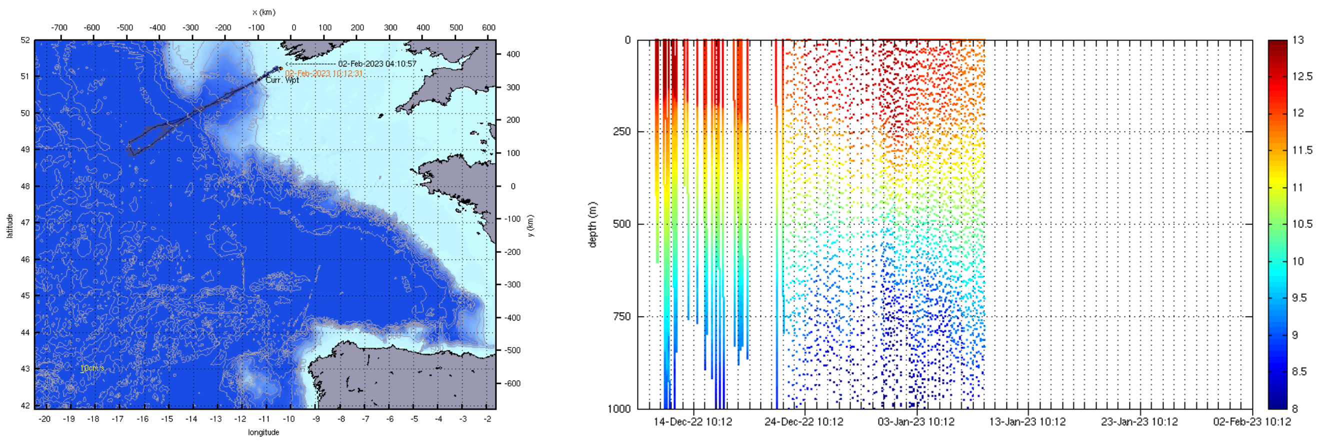

Figure 4. Marine Institute glider Glider *Aisling na mara* trajectory (left) and water temperature data (right) during the iFADO PAAnoramic mission as depicted in EGO-network website. EGO-Network link for this mission.

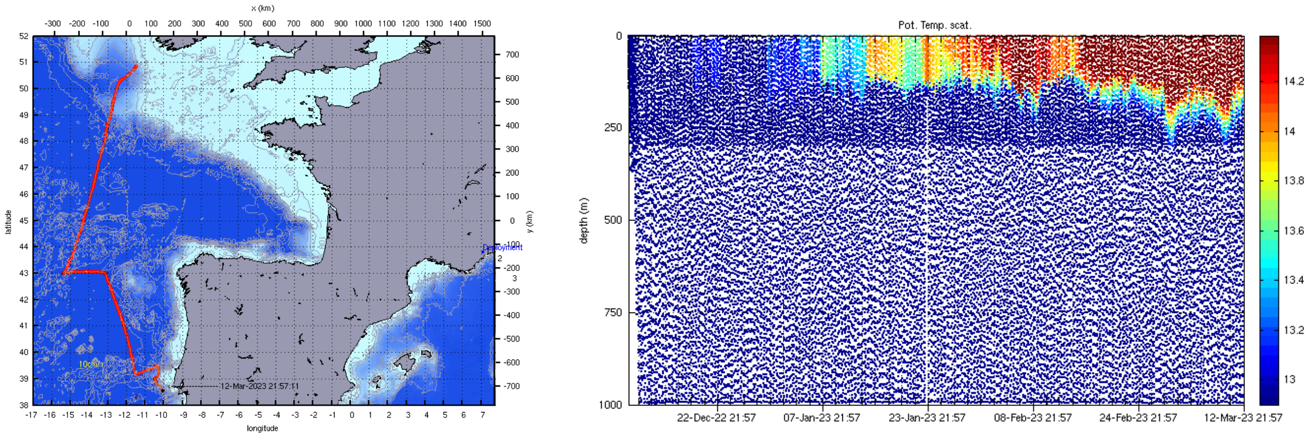

Figure 5. National Oceanographic Centre glider trajectory (left) and water temperature data (right) during the iFADO PAAnoramic mission as depicted in EGO-network website. The vehicle used for this leg was Sg152 operated by Cyprus Subsea Consulting & Services. All the information can be found at this link EGO-Network link for this mission.

References

Harscoat V, Pouliquen S (2016). Data Management Handbook. H2020 AtlantOS project D7.4. Available at https://www.atlantos-h2020.eu/download/7.4-Data-Management-Handbook.pdf. [Last access: 20 February 2023].

Unidata, NcML Overview 2023 [Online]. Available: https://docs.unidata.ucar.edu/netcdf-java/current/userguide/ncml_overview.html [Last access: 20 February 2023].