# Connecting ocean physics near real time data streams to the European Marine Observation and Data network

# Index

# Introduction

EMODnet - the European Marine Observation and Data network – is a long term marine data initiative from the European Commission Directorate-General for Maritime Affairs and Fisheries (DG MARE) involving and networking more than 150 organizations for assembling marine data, products, and metadata. The data infrastructure has been developed through a stepwise approach in 3 major phases by running 8 thematic portals, 6 regional check points and a Data Ingestion facility.

EMODnet Physics (www.emodnet-physics.eu), one of the thematic portals, is developing a combined array of services and functionalities such as facility for viewing and downloading, dashboard reporting and machine-to-machine communication services, to obtain, free of charge data, meta-data and data products on the physical conditions of the ocean from many different distributed data sets.

The EMODnet Data Ingestion portal seeks to identify and to reach out to other potential providers in order to make their data sets also part of the total offer. It aims at streamlining the data ingestion process so that data holders from public and private sectors that are not yet connected to the existing marine data management infrastructures can easily release their data for safekeeping and subsequent distribution through EMODnet. This will enrich the total offer for all types of users and conform to the EMODnet motto 'collect data once and use it many times'.

The scope of this EMODnet Physics – EMODnet Data Ingestion joint document is to present the data stream model considering the role of the different players that are involved in the data production and dissemination pipeline.

It takes into consideration the role of the data providers, the regional ocean observing systems, the EuroGOOS, and the major European infrastructures for data management and dissemination (i.e. the Copernicus Marine Service Environmental Monitoring System and its in situ thematic assembly center, the SeaDataNet network of the National Oceanographic Data Centers and the EMODnet Physics)

The document is aiming at being an introductionary guide to data providers that are willing to share and make their data available into the EMODnet (Physics) portal.

This overview is following one key recommendation from the annual EuroGOOS meeting in 2010, i.e. it is essential to meet the following needs:

- *provision of easy access to data through standard generic tools, easy means of using the data without having to be concerned about data processing and who processes them, and that adequate metadata are available to describe how the data were processed.*

## Data stream, roles and data management infrastructures

Institutions are in charge of the management of the observing system and its first level of data processing for their own applications. An institution collects, controls and distributes data according to its own rules but the development of Operational Oceanography involves major investments in infrastructures, including observing systems and high performance computing hardware, as well as human re-sources with appropriate training. Such investments are difficult to be made by a single country and thus active cooperation has always been key for the development of Global Ocean Observation System. It is particularly true when it comes to open ocean systems either at regional or wider scale.

In order to tackle this critical issue, since 1994, EuroGOOS is coordinating the development and operation of (European) regional operational systems. Five systems are at present part of EuroGOOS: the Arctic (Arctic ROOS), the Baltic (BOOS), the North West Shelf (NOOS), the Ireland-Biscay-Iberian area (IBI-ROOS) and the Mediterranean (MONGOOS). EuroGOOS also contribute the Global Ocean System as one GRA of GOOS and in partnership with JCOMM.

These regional assemblies are the key structures in which it is possible to discuss to promote active cooperation at different levels in order to maximize the efficiency of national resources and investments in operational oceanography. This is done via specific and thematic working groups that collect and express the best expertise on specific fields.

The ocean data management and exchange process within EuroGOOS are intended to reduce duplication of effort among agencies, improve quality and reduce costs related to geographic information, thus making oceanographic data more accessible to the public and helping to establish key partnerships to increase data availability. Recent EU marine data infrastructures and EU Programs are widely based on EuroGOOS and ROOSs achievements.

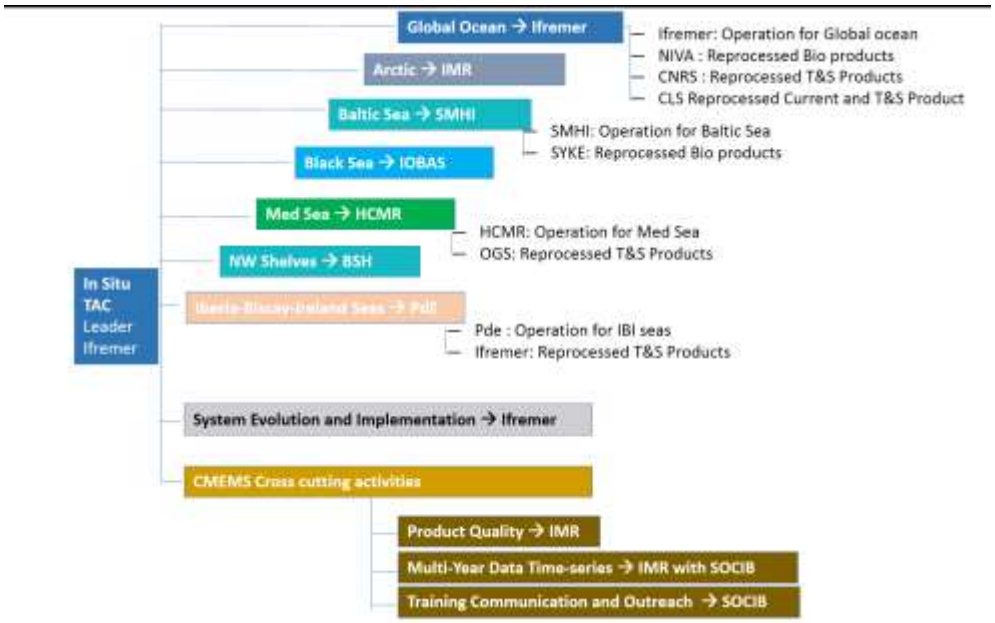
The major infrastructures are the Copernicus Marine Environmental Monitoring Service (CMEMS), the SeaDataNet (SDN) network of National Oceanographic Data Centres and the European Marine Observation and Data Network (EMODnet) - the Physics lot.

The Copernicus Marine Service has been designed to respond to issues emerging in the environmental, business and scientific sectors. Using information from both satellite and in situ observations, it provides state-of-the-art analyses and forecasts daily, which offer an unprecedented capability to observe, understand and anticipate marine environment events. The CMEMS In Situ Thematic Assembly Center (INSTAC) was designed and developed on JCOMM and the EuroGOOS ROOSs experience and expertise, which was further developed during the MyOcean projects. MyOcean enabled to run a demonstration pre-operational service for 6 years that is now fully integrated and constituting the CMEMS INSTAC.

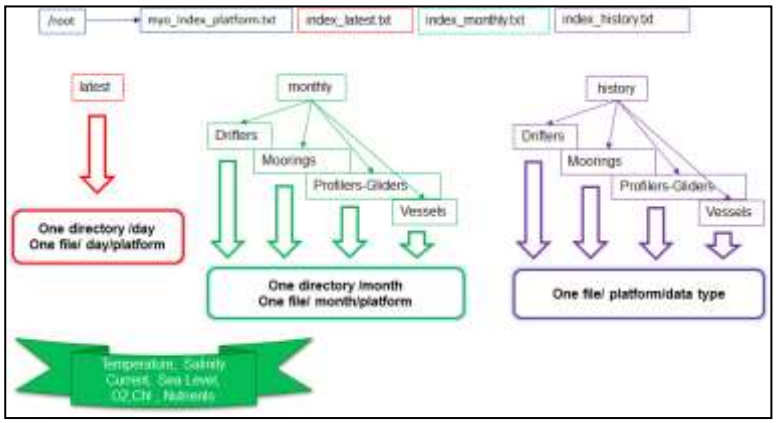
For each EuroGOOS Region there is a Regional Data Assembly Centre (RDAC) operated jointly with the CMEMS INSTAC and working closely with organisations operating monitoring stations. In this

federative infrastructure, the quality of the products delivered to users must be equivalent wherever the data are processed: each RDAC is responsible for assembling data provided by institutions and provides a unique data access point to bundle available data into an integrated dataset for validation and distribution (whereby validation is following common EuroGOOS DATAMEQ –CMEMS- EMODnet harmonized procedures). Each RDAC validates the dataset consistency in their area of responsibility, typology of data and typology of parameter.

The scheme shows the INSTAC hierarchical architecture.




Routinely (e.g: every hour), each RDAC distributes all its new data on its regional portal. Files (i.e. NetCDF files) are organized in folders as described in the following:




During the INSTAC operational activities, quality control is performed automatically on the data that is made available in real-time and near real-time, yearly scientific assessment performed on the latest 30 years of data ( 60 years for T&S and Global scale).

Further to this, some networks are organized with **Global Data Assembly Centre** (GDAC)

components. It is designed for a global observation network such as Argo, OceanSITES, EGO for Gliders, etc. The GDAC aggregates data and metadata provided by Network DACs, in RT (Real Time) and DM (Delayed Mode).

A further validation and quality control of the data take place when the observation data are passed to data centres that also take care of adding more extensive metadata and ensuring long-term storage and stewardship. These data centres, of which many are National Oceanographic Data Centres (NODCs), work together in the pan-European SeaDataNet network to develop and promote marine data management standards as well as tools and services for applying these standards for good marine data management. The network also has built up and is operating and further developing the SeaDataNet infrastructure for managing, processing, discovery and giving access to a wide range of marine and ocean data, covering physics, but also chemistry, geophysics, geology, and biology data sets. This way the SeaDataNet network is an important source of data for EMODnet and CMEMS.

The European Marine Observation and Data Network (EMODnet) is a long-term programme to deliver a marine observation infrastructure offering effective support to the marine and maritime economy whilst supporting environmental protection needs. Currently EMODnet is organized in seven sub-portals providing access to marine data from the following themes: bathymetry, geology, physics, chemistry, biology, seabed habitats and human activities. The EMODnet lots are developing high quality data products with a harmonized time and space resolution across the Europe.

Despites the others, EMODnet Physics is also making available and accessible in situ near real time and historical validated marine and ocean data as monitored by in situ fixed and moving platforms such as fixed stations, mooring buoys, tide gauges, surface drifters, ferryboxes, argo floats, gliders, HF radars etc.

Each platform may provide one or more physical parameters of the sea, namely sea surface temperature, temperature in the water body column, sea surface salinity, sea salinity in the water column, sea surface currents, sea level, wave direction, wave peak, waves frequency, atmospheric pressure at sea level, sea turbidity, chlorophyll(a), etc.

The variety of the available platforms, as well as the variety of their instruments (and accuracy) is very diverse within the different ocean regions.

Each platform and each parameter requires a specific data management flow and may go under /or different levels of qualification (automatic, operator and scientist based) that have been agreed at European and/or international level. These actions are crucial before combining data from different sources: it is crucial that the data are comparable and compatible to avoid mistakes in analyses and interpretation.

In this very well organized and coordinated framework, the establishment of new data stream and data flow has to consider and serve described principles and thus has to fill and be available into the described relevant EU infrastructures.

## Data conventions

The very first principle to serve is cross platform coherence. It requires a minimum and essential set of applicable recommendations relying on existing international standards and, in turn, it facilitates data discovery for users and data integration. Dataset have be integrated with essential set of metadata covering the platform identifier, the parameters, the provider and the quality of the data.

| Platform | Parameters | Provider | Data quality |
|---|---|---|---|
| The platform should have a unique identifier. It should be either the WMO code, or ICES code for ships. If the platforms does not follow in these registries, a regional/thematic id has to be identified. | SDN controlled vocabularies managed by NERC/BODC are recommended for the physical parameters: P01 (parameter), P07 (CF variable), P06 (units) from | The European Directory of Marine Organizations (EDMO) is a catalogue of the oceanographic institutes. It provides a controlled registry for identifying the Institution providing the data | Data quality information (Quality Flag) has to be attached to the data. |

## Data quality

Quality Management usually consists of the two phases: Quality Assurance (QA) and Quality Control (QC). While the QA is the set of procedures and protocols to be applied before and during the data set creation, the QC is the inspection of the dataset for conformance to the data product specification. While doing the QC, datasets are assigned a quality check flag (QF).

The QC/QF of data is an essential component of oceanographic data management: it tells the users of the data how data was gathered, how it was checked, processed, what algorithms have been used, what errors were found, and how the errors have been corrected or flagged.

The QC/QF enables the possibility to combine, integrate, re-use data from different sources[1]

EMODnet Physics "bibliography" page (http://www.emodnet-physics.eu/portal/bibliography) hosts the list and links to the available best practices on QA, QC/QF.

For more exhaustive description and references on QA methodologies it is recommended to consult the British Columbia Field Sampling Manual[2] and, in general, the QARTOD Manuals for methodologies and references on QA, QC/QF[3].

Quality control procedures for real time, delayed mode and reprocessed data are collected into an

[1] http://library.oceanteacher.org/OTMediawiki/index.php/Marine_Data_Quality_Control
[2] http://www.env.gov.bc.ca/epd/wamr/labsys/field_man_pdfs/fld_man_03.pdf
[3] http://www.ioos.noaa.gov/qartod/welcome.html

evolving Wiki4 (https://150.145.136.231:8081/WordPress/).

The JERICO-Next Deliverable 2.2[5] offers an exhaustive overview on sensors and sensor calibration.

## Data transport formats

The data transport files format for the NRT data exchange is an implementation of NetCDF OceanSITES format. If the user can already make this data format available, the RDAC/GDAC have only to uptake this new data and update the data index. If the provider is willing to provide data by means of a different data format (e.g. txt) or means (e.g. API SOAP/REST, OGC SOS, etc), the data center has is going to create an adapter to this data stream and create the netCDF.

## Data server, data transfer

### FTP

FTP portal are common means for distributing data. They are based on a quite consolidated and robust data transfer method and can be easily implemented, maintained and monitored. FTP portals are working properly when the amount of data to be hosted/transferred is not too big and complex. E.g. FTP portals are working properly for platforms creating time series of data.

### THREDDS

In situ gridded data like HF Radar data is big data (2D data sampled at high frequency) and has to be managed according its peculiarity and complexity. The up to date and most adopted solution is based on THREDDS services. Thematic Real-time Environmental Distributed Data Services (THREDDS) that consists of two main building blocks: the THREDDS Data Server (TDS) and the Common Data Model (CDM) / netCDF-Java library. The TDS is open source and runs inside the open source Tomcat Servlet container. The TDS provides catalog, metadata, and data access services for scientific data. Every TDS publishes THREDDS catalogs that advertise the datasets and services it makes available. THREDDS catalogs are XML documents that list datasets and the data access services available for the datasets. Catalogs may contain metadata to document details about the datasets. TDS configuration files provide the TDS with information about which datasets and data collections are available and what services are provided for the datasets. The available remote data access protocols include OPeNDAP, OGC WCS, OGC WMS, and HTTP. It has to be noticed that the ncISO service allows THREDDS catalogs to be translated into ISO metadata records.

The CDM provides data access through the netCDF-Java API to a variety of data formats (e.g., netCDF, HDF, GRIB). Layered above the basic data access, the CDM uses the metadata contained in datasets to provide a higher-level interface to geoscience specific features of datasets, in particular, providing geolocation and data subsetting in coordinate space.
The TDS uses the CDM/netCDF-Java to read datasets in various formats. The CDM also provides the foundation for all the services made available through the TDS.

[4] The Wiki is a CNR ISMAR (SP Italy) and ETT initiative.
[5] http://www.jerico-ri.eu/download/jerico-next-deliverables/JERICO-NEXT-Deliverable-2.2.pdf

A pluggable framework allows other developers to add readers for their own specialized formats. The CDM also provides standard APIs for georeferencing coordinate systems, and specialized queries for scientific feature types like Grid, Point, and Radial datasets, and so it represents the best suitable available technology to manage in situ gridded data products.

An exhaustive guide is available at:
http://www.unidata.ucar.edu/software/thredds/current/tds/reference/ChecklistReference.html

### ERDDAP

ERDDAP is an Apache based data server that offers an easy and consistent way to download subsets of gridded and tabular scientific datasets in common file formats and make graphs and maps. the peculiarity of ERDDAP is that it unifies the different types of data servers so you have a consistent way to get the data you want, in the format you want. In particular, ERDDAP reformats the request into the format required (.html table, ESRI .asc and .csv, Google Earth .kml, OPeNDAP binary, .mat, .nc, ODV .txt, .csv, .tsv, .json, and .xhtml) by the remote server, sends the request to the remote server, gets the data, reformats the data into the format that you requested, and sends the data to requester.
A very detailed introduction to ERDDAP is available at:
https://coastwatch.pfeg.noaa.gov/erddap/information.html

### Sensor Web Enablement

The OGC Sensor Web Enablement (OGC-SWE) defines standards to exploit Web-connected sensors and sensor systems, such as oceanographic instruments [OGC 06-021r4]. In particular, the OGC-SWE enables the discovery, exchange, and processing of sensor observations and tasking of sensor systems by means of:

- 3 standards XML encodings (SensorML, O&M, TML)
- 4 standards web service interfaces (SOS, SAS, SPS, WNS)

Some of these standards are specifically designed to enhance data interoperability and discovery and so they are tools that providers can implement to make their data available and accessible.
The components that enable data discovery and exchange are:

- Sensor Model Language (SensorML) – Standard conceptual model and XML schema to describe sensors, systems, and processes. SensorML provides the information needed for the discovery of sensors, location of sensor observations, configuration of sensor networks, processing of low-level sensor observations, and listing of ”task-able” processes
- Observations and Measurements (O&M) – Standard conceptual model and XML schema to encode observations and measurements. O&M defines an “observation” as an event whose result is an estimate of the value of some property of a feature of interest, obtained using a specified procedure. A sensor, channel, or systems of sensors could all be treated as a procedure. Data interoperability between instruments can be achieved with the O&M standard.
- Transducer Markup Language (TML) – Conceptual model and XML schema to describe transducers and real-time streaming of data to and from sensor systems.

- Sensor Observation Service (SOS) - Standard web service interface for requesting, registering, filtering, and retrieving observations and sensor system information. SOS is the intermediary between a client and an observation repository or near real-time sensor channel.

The full list of OGC standards is available at http://www.opengeospatial.org/standards

The OGC SWE also supports and enables real time data streams: EMODnet Data Ingestion is developing a specific pilot to show off the concept and facilitate the real time ingestion of physical parameters into the EMODnet Physics. This pilot makes use of developments in a series of EU-projects. A first draft of the SWE profiles is now published under https://odip.github.io/MarineProfilesForSWE/. This site includes a story that narrates how projects, people, technologies and vocabularies were brought together to formulate meaningful and semantically rich profiles for the marine domain. The related EU-projects that have funded this effort are listed under the above mentioned URL.

## Implementation principles

Although the in situ data management structure is quite complex, it is well coordinated at European level (EuroGOOS – EMODnet Physics – CMEMS - SDN). As introduced in the previous sections, it is based on a hierarchical regional infrastructure that facilitates the management and integration of any potential data provider or data source.
The implementation works according to a couple of simple and very effective rules:

- If the data provider can set up the data flow according the defined standards, the regional coordinator only has to link and include the new catalogue and data stream
- If the data provider cannot setup the data flow (because of lack of experience, technical capacity etc), the regional coordinator has to work on harvesting the data from the provider, harmonize and format these data and make them available from the regional catalogue.

More specifically the Regional coordinator is going to implement the following functions:

Data acquisition:

- Data are collected through direct links with the institutions
    - Direct connection is established (see data server, data transfer section) between the RAC and the data provider
    - Information is provided about the required metadata that should be supplied together with data (*ex. station position, date, frequency of measurement, platform name, depth of each sensor, contact person, PI, etc.*)
      ***→ the RDAC staff collects the needed information to fill up the required metadata***
    - Guidance is also provided on how the required daily and monthly files should be created.
    - Information exchanged about the QC procedures

**QC Flag Scale**

| code | Meaning |
|---|---|
| 0 | no QC was performed |
| 1 | good data |
| 2 | probably good data |
| 3 | bad data, but correctable |
| 4 | bad data |
| 5 | value changed |
| 6 | below detection limit |
| 7 | in excess of quoted value |
| 8 | interpolated value |
| 9 | missing value |
| A | Incomplete information |

mandatory for automated QC

    - Data are provided in the originator's native format, no need for conversion to NetCDF.
      ***→ This procedure is performed by the RDAC staff***
- Data are converted in a unique format (netcdf)
- Extended guidance for new partners in order to provide all the necessary information

Quality Control:

- Apply automatic quality control procedures on each parameter, elaborated in coherence with international agreement (in particular SeaDataNet).
- Procedures applied after agreement with the data originators in order to avoid conflicts and effort duplications.

Validation/Assessment:

- Assess the consistency of the data over a period of time in an area. The aim is to detect possible incoherencies with nearby data that could not be detected by automatic QC.

Then data can be further validated and stored and saved for long term stewardship (at national or regional or European level - according the provider tech and infrastructure capacity).

# Contact points for connecting new operational data sources and new data streams

As described previously, according the typology of platform some steps may change, the following table provides different contacts within the EMODnet Physics network to try and support at the best the provider needs.

| | name | | | email |
|---|---|---|---|---|
| EMODnet Physics | Antonio Novellino | ETT | Coordinator | antonio.novellino@ettsolutions.com |
| EuroGOOS | Patrick Gorringe | EuroGOOS | Deputy | Patrick.gorringe@smhi.se |
| EuroGOOS DATAMEQ | Sylvie Pouliquen | IFREMER | Chair | Sylvie.Pouliquen@ifremer.fr |

Platform type – platform network

| | name | | | email |
|---|---|---|---|---|
| Tide gauge | Begoña Pérez | Puertos del Estado | Task Team Chair | bego@puertos.es |
| Glider | Pierre Testor, | Université Pierre et Marie CURIE | Task Team Chair | testor@locean-ipsl.upmc.fr |
| HF radar | Julien Mader | AZTI | Task Team Chair | jmader@azti.es |
| Animal borne Instruments | Lars Boehme | University of St Andrews | Task Team Chair | lb284@st-andrews.ac.uk |
| Ferrybox | Wilhelm Petersen Franciscus Colijn | HZG | Task Team | wilhelm.petersen@hzg.de; franciscus.colijn@hzg.de |
| EMSO ERIC + Fixed Station | Richard Lampitt | NOC | Task Team Chair | r.lampitt@noc.ac.uk |
| ARGO and profiling buoys | Sylvie Pouliquen | IFREMER | Task Team Chair | Sylvie.Pouliquen@ifremer.fr; |
| others | Patrick Gorringe – Antonio Novellino | | | Patrick.gorringe@smhi.se; antonio.novellino@ettsolutions.com |

Regional ROOS Data Manager - CMEMS INSTAC RDAC

| | name | | email |
|---|---|---|---|
| Arctic | Sjur Ringheim | IMR | cmems-service@imr.no |
| Baltic | Johanna Linders | SMHI | cmems-service@smhi.se |
| IBI | Marta de Alfonso | Puertos del Estado | cmems-service@puertos.es |
| North West Shelf | Susanne Tamm | BSH | cmems-service@bsh.de |
| Mediterranean Sea | Leonidas Perivoliotis | HCMR | cmems-service@hcmr.gr |
| Global Oceans | Thierry Carval | IFREMER | cmems-service@ifremer.fr |
| Black Sea | Veselka Marinova | IOBAS | cmems-service@io-bas.bg |

Not EU data

| | | |
|---|---|---|
| EMODnet Physics | Antonio Novellino | antonio.novellino@ettsolutions.com |

**Note**: In addition arrangements can be made with a SeaDataNet data centre for further validation of the collected datasets and inclusion in the data management infrastructure for long term

stewardship. Alternatively the data provider can decide to ingest the datasets by means of the Data Submission service at this portal whereby it will be received by a SeaDataNet data centre for further processing.