www.nautilos-h2020.eu

Organisation ETT SpA

D8.4

Design of Thematic Assembly Centre for innovative parameters

Date: 30th September 2021
Doc. Version: 1.3
10.5281/zenodo.7211817

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101000825 (NAUTILOS). This output reflects only the author's view and the European Union cannot be held responsible for any use that may be made of the information contained therein.

Document Control Information

Settings	Value
Deliverable Title	Design of Thematic Assembly centre for innovative parameters
Work Package Title	Data Management
Deliverable number	D8.4
Description	NAUTILOS is developing innovative sensors, some are going to improve the actual sensor accuracy and resolution, some others are tackling the acquisition of new parameters. The deliverable describes common methods for parameter-platform management, from collection (and procedures for data quality insurance) to dissemination. Dissemination considers standards and methods to facilitate interoperability towards other European Infrastructure such as EMODnet, CMEMS, SeaDataNet etc. This document defines the methodology and the structure to implement a Thematic data Assembly centre (TAC) for parameters and data streams that are not yet consolidated under a European or International Data/Thematic assembly centre nor under one of the major European data infrastructures.
Lead Beneficiary	ETT SpA
Lead Authors	Antonio Novellino
Contributors	Federica Colombo
Submitted by	Federica Colombo
Doc. Version (Revision number)	1.3
Sensitivity (Security):	Public
Date:	30th September 2021
DOI	10.5281/zenodo.7211817

Document Approver(s) and Reviewer(s):

NOTE: All Approvers are required. Records of each approver must be maintained. All Reviewers in the list are considered required unless explicitly listed as Optional.

Name	Role	Action	Date
Gabriele Pieri	Coordinator/WP1 leader	*Review*	3rd August 2021
Andy Smerdon		*Review*	4th August 2021

Document history:

The Document Author is authorized to make the following types of changes to the document without requiring that the document be re-approved:

- Editorial, formatting, and spelling
- Clarification

To request a change to this document, contact the Document Author or Owner.
Changes to this document are summarized in the following table in reverse chronological order (latest version first).

Revision	Date	Created by	Short Description of Changes

Configuration Management: Document Location

The latest version of this controlled document is stored in <location>.

Nature of the deliverable		
R	Report	
DEC	Websites, patents, filing, etc.	
DEM	Demonstrator	
O	Other	X

Dissemination level		
PU	Public	X
CO	Confidential, only for members of the consortium (including the Commission Services)	

ACKNOWLEDGEMENT

This report forms part of the deliverables from the NAUTILOS project which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101000825. The Community is not responsible for any use that might be made of the content of this publication.

NAUTILOS - New Approach to Underwater Technologies for Innovative, Low-cost Ocean observation is an H2020 project funded under the Future of Seas and Oceans Flagship Initiative, coordinated by the National Research Council of Italy (CNR, Consiglio Nazionale delle Ricerche). It brings together a group of 21 entities from 11 European countries with multidisciplinary expertise ranging from ocean instrumentation development and integration, ocean sensing and sampling instrumentation, data processing, modelling and control, operational oceanography and biology and ecosystems and biogeochemistry such, water and climate change science, technological marine applications and research infrastructures.

NAUTILOS will fill-in marine observation and modelling gaps for chemical, biological and deep ocean physics variables through the development of a new generation of cost-effective sensors and samplers, the integration of the aforementioned technologies within observing platforms and their deployment in large-scale demonstrations in European seas. The fundamental aim of the project will be to complement and expand current European observation tools and services, to obtain a collection of data at a much higher spatial resolution, temporal regularity and length than currently available at the European scale, and to further enable and democratise the monitoring of the marine environment to both traditional and non-traditional data users.

NAUTILOS is one of two projects included in the EU´s efforts to support the European Strategy for Plastics in a Circular Economy by supporting the demonstration of new and innovative technologies to measure the Essential Ocean Variables (EOV).

More information on the project can be found at: http://www.nautilos-project.eu.

COPYRIGHT

TABLE OF CONTENTS

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Data assembly centres (DACs) play an active role in ocean data management. They collect data, harmonize data in terms of format and metadata, provide quality-controlled data, making them discoverable together with their metadata.

This report presents an analysis of the best practices and recommendations adopted for the management of the parameters analysed in the NAUTILOS project. As new sensors and samplers are being developed within this project, this report also provides recommendations and data management structures that can easily interoperate with and be included in already established European data infrastructures.

We analysed existing Thematic Assembly Centres, highlighting common elements and the gaps to be designed to fulfil NAUTILOS scope. The analysis shows that DACs are well established for most of the platforms deployed throughout the NAUTILOS project and in a few cases, i.e. the NAUTILOS innovative sensors, the need is to work towards the adoption of some standardised model to facilitate interoperability.

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS AND ABBREVIATIONS

Abbreviation	Definition
AniBOS	Animal borne ocean sensors
CMEMS	Copernicus Marine Environment Monitoring Service
DAC	Data Assembly Centre
EOV	Essential Ocean Variable
FB	Ferry box
FTP	File transfer protocol
GBIF	Global Biodiversity Information Facility
GDAC	Global data assembly centre
GOOS	Global ocean observing system
GOSUD	Global Ocean Surface Underway Data
IOC	International Oceanographic Commission
IODE	International Oceanographic Data Exchange
IPT	Integrated publishing toolkit
JCOMMOPS	Joint Technical Commission for Oceanography and Marine Meteorology in situ Observing Platform Support Centre
MSFD	Marine Strategy Framework Directive
OA	Ocean acidification
OBIS	Ocean Biodiversity Information System
PI	Principal investigator
ROOS	Regional ocean observing system
SOCAT	Southern Ocean Carbon Atlas
SOOP	Ships of Opportunity Programme

TAC	Thematic Assembly Centre
VOS	Voluntary Observing Ships
WMO	World Maritime Organisation

1. INTRODUCTION

NAUTILOS is using seafloor landers, Argo floats, animal-borne instruments, ferry boxes, ships of opportunities and fishing vessels to study physical, chemical and biological processes.

Physical ocean data represent the "physical properties and dynamic processes of the oceans," including how the ocean interacts with the atmosphere, ocean temperature, currents, coastal dynamics and more. In this field, NAUTILOS targets temperature and salinity data.

Chemical ocean data relates to the chemical makeup, processes and cycles of ocean waters as well as how seawater interacts with the atmosphere and the seafloor. Biogeochemical data relates to the cycling of nutrients from the biotic environment or biosphere (i.e., living organisms) to the abiotic environment, which includes the atmosphere, lithosphere and hydrosphere, and vice versa. Some of the chemical parameters needed to understand this process include oxygen, nutrients, inorganic carbon, particulate matter, nitrous oxide, stable carbon isotopes, dissolved organic carbon and ocean colour. NAUTILOS is addressing inorganic carbon, nutrients, and oxygen. Micro- and nano-plastics also fall into this category while suspended particulate matter, ocean colour are in between physical and chemical parameters.

Biological ocean data applies specifically to marine organisms and how they interact with the ocean environment. Datasets usually record the abundance, composition and/or behaviour of marine life. NAUTILOS is addressing phytoplankton biomass and diversity, zooplankton biomass and diversity, turtles, marine birds, marine mammals abundance and distribution, live coral, seagrass cover, microbial biomass and diversity and invertebrate abundance and distribution. Ocean sound may fall under physical (underwater noise) and under biological data (sound produced by marine animals).

These datasets are of paramount importance as they can contribute to the global ocean observing system (GOOS) under the framework of ocean observation (FOO) (Lindstrom et al., 2012). Therefore, a preliminary requirement for NAUTILOS is to make these data FAIR and accessible to third parties. While developing interoperability interfaces towards established European infrastructures (e.g. CMEMS INS TAC, EMODnet Physics, EMODnet Chemistry, SeaDataNet, etc.), one goal for the NAUTILOS project is to investigate Thematic Data Assembly Centres' workflow, highlighting common elements to be adopted in the project data management, and potential gaps to be addressed under the NAUTILOS scope.

The analysis shows that DACs are well established for most of the platforms deployed throughout the NAUTILOS project and in a few cases, i.e. the NAUTILOS innovative sensors, working towards the adoption of some standardised model is needed to facilitate interoperability. This report presents the results and the next steps for NAUTILOS data management.

2. OCEAN OBSERVATION AND DATA CENTRES

The Framework for Ocean Observation (FOO) defines the methodology for the ocean observing community toward the establishment of an integrated, sustained ocean observing system. The FOO collaborative structure implements best practices for essential ocean variables data collection. Pillar components of this "system of system" infrastructure are data management and data interoperability standards.

Data assembly centres (DAC) and thematic data processing centres are essential elements of this operational oceanography infrastructure. These centres have been developed nationally (National Oceanographic Data Centre), regionally (Regional Thematic DAC) or globally (Global DAC). They maintain, update, and provide access to marine environmental and ecosystem data, metadata and data products.

According to the IODE (International Oceanographic Data Exchange) the role of a (thematic) data center (TAC) is to collect and harmonise data, often quality control procedures apply quality flags, and sometimes a data centre may process data to facilitate the end-user uptake and use of the data for monitoring, modelling or downstream service development.

As a TAC may have to apply some processing (e.g. integrity check, quality check, format conversion), it usually holds an inventory of the data sources.

Operational oceanography needs two types of data: near real-time data required for daily and weekly forecasting activities, and delayed-mode data that are subject to greater quality control. These data are particularly valuable for reanalysis work and to assist seasonal forecasting and long-term climate monitoring and prediction.

Some platforms, such as gliders or research cruises, may not transmit (all) the data in real time to the TAC. In this case, we may have an operational stream that includes a subset of parameters delivered in near real time, while the complete dataset is recovered at the end of the mission (i.e. delayed mode). Other delayed mode observations derive from the reprocessing of near real time observations that undergo more exhaustive and advanced quality check procedures (see next paragraphs for details).

IODE terms of reference for a GDAC

A data assembly centre has to:

- receive and assemble marine meteorological and/or oceanographic data (real or delayed-mode) and metadata from the appropriate data streams and check these are consistent;
- identify duplicates and if possible resolve by keeping the best copy of a dataset;
- make sure that the data are quality controlled according to the international standards and methods established by IODE, WMO or JCOMM as appropriate;
- provide feedback to the sources of data regarding quality issues;
- make data accessible through IODE/ODP;
- make discovery metadata available to IODE/ODP;
- forward data and metadata to the appropriate CMOC(s) in agreed format(s) within defined timescales;
- contribute to WMO and IOC Applications by collecting and processing worldwide marine-meteorological and oceanographic data and metadata documented in appropriate WMO and IOC publications;

- report to the IODE and JCOMM Committees on data management status and activities

During the past 50 years a number of DACs/TACs and GDACs have been consolidated. in this report we briefly describe the key elements of those relevant for NAUTILOS, namely ARGO, OceanSITES, AniBOS, OceanGlider, and research vessels.

- a TAC is, in general, set up to provide a single-entry point to data processed in national centres, applying commonly defined quality control procedures at all steps of data processing.
- a TAC usually manages two data streams:
    - a real-time data (distributed in less than 24 hours) that are quality-controlled and flagged using an automated procedure (data are free from gross errors and may be corrected in real time when the correction is known)
    - a delayed mode, data are produced later (over 1 year) and require the control and validation by a scientific expert.
- a TAC makes data available in a common transport format and data model (and netCDF and Climate and Forecast convention have become the *de facto* standard).
    - data model includes information about data integrity and data quality (according to standardised quality check and quality flag procedures)
- TACs/DACs may have several federated nodes that may be synchronised or may feed a further aggregation level, e.g globally (GDACs).
- a TAC implements open data access and open file transfer protocols (FTP, http, OpenDAP, ERDDAP)

These recommendations can apply to operational data mainly falling under the physics domain and (partially) the chemistry domain. For biological data it is important to keep in mind the main European and global initiatives on data integration and work towards making NAUTILOS outputs compatible with those initiatives. We briefly present these main initiatives in the following sections.

2.1. ARGO FLOATS

The Argo network is a global array of more than 3,500 autonomous instruments, deployed in the world ocean, reporting subsurface ocean properties. This network has revolutionised the distribution of ocean data within the research and operational communities (Roemmich et al., 2009).

Argo led to a new paradigm for oceanography, making all data and data products freely available in real time. This is achieved through two GDACs, which are synchronised and act as backup of each other: the Coriolis Data centre in France and the US Navy's Fleet Numerical Meteorology and Oceanography centre (FNMOC).

The data management issue for Argo was to set up an information system able to provide a single entry point to data processed in national centres, applying commonly-defined quality control procedures at all steps of data processing. Two data streams have been identified: first, a real-time data stream where data are free from gross errors and may be corrected in real time when the correction is known, and second, a data stream that operates in a delayed mode, where data profiles have been subjected to detailed scrutiny by oceanographic experts (Wong et al., 2020).

Data is available in a common netCDF format and can be downloaded by e.g. File Transfer Protocol (FTP), web, and ERDDAP. The two Argo GDACs receive data that have been processed by 11 national DACs. Each float is allocated to a specific DAC. Data holdings at the two Argo GDACs are synchronised once per day.

For floats equipped with biogeochemical sensors, vertical profile data are stored in two separate profile data files. A file checker checks the format and content consistency of these data files before they are admitted into the global data holdings (Ignaszewski, 2018).

This architecture has proven to be efficient, robust, able to serve both operational and research communities, and sustainable in the long term by relying on professional data centres. This model was adopted by other international programmes, such as Global Oceanographic Surface Underway Data (GOSUD) and OceanSITES (Deep Ocean Eulerian observatories), which have both DACs and GDACs and have extended the Argo netCDF format to handle their data.

Lately, the Biogeochemical-Argo programme developed a global network of biogeochemical sensors on Argo profiling floats, including chlorophyll a fluorescence sensors.

Biogeochemical Argo data management follows Argo data management rules. Basically data are made publically available in real-time and delayed mode.

The Biogeochemical-Argo data management group develops and implements the procedures for quality control of the core biogeochemical variables (https://biogeochemical-argo.org/data-management.php).

2.2. OceanSITES (Fixed Moorings)

OceanSITES is a worldwide system of long-term, deepwater reference stations measuring dozens of variables and monitoring the full depth of the ocean, from air-sea interface down to a depth of 5,000 metres.

OceanSITES operates long-term buoy and ship stations that measure many aspects of the ocean's surface and depth profile using advanced sensors (Figure 1), and make data available in real time via satellite telemetry and regional and global DACs.

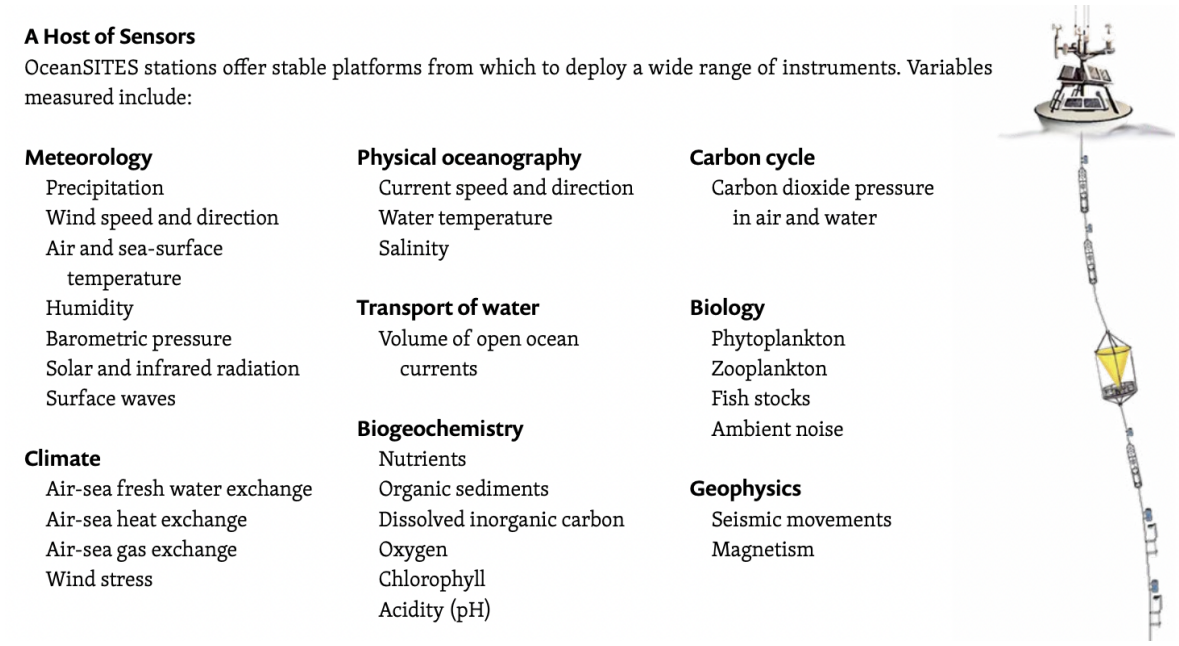

Figure 1 OceanSITES

OceanSITES coordinates the open ocean time series activities for the Global Ocean Observing System (https://www. goosocean.org). OceanSITES data are provided in NetCDF data format with supporting metadata. These data are served by two GDACs, one hosted by Coriolis at Ifremer in Brest, France and one at the National Data Buoy Centre (NDBC) in the United States.

2.3. Gliders and OceanGliders

Gliders are autonomous underwater vehicles that perform saw-tooth trajectories from the surface to depths of 1,000 m, along programmable routes. They move up(down)ward thanks to a buoyancy engine and they achieve forward speeds of up to 40 km/day thanks to wings and rudders. They can be operated for months and over thousands of km before they have to be recovered. Gliders record Essential Oceanic Variables (physical and biogeochemical) at high resolution during the dives and transmit these data in near real time to land via satellite when at surface (every few hours).

OceanGliders programme has recently been recognised by the OOPC (Ocean Observations Panel for Climate), and the IOC and WMO's JCOMM OCG (Joint Commission on Oceanography and Marine Meteorology Observation Coordination Group), and is now engaging in the GOOS through these mechanisms. The European component of the OcenGliders is Everyone's Gliding Observatories (EGO) which data management is based on what has been designed for the Argo and OceanSites data management.

The glider's data flow is carried out through four organisational units: glider operators, principal investigators (PIs), data assembly centres (DACs) and global DACs (GDACs). The PI, typically a scientist at a research institution, maintains the observing platform and the sensors that deliver the data. They are responsible for providing the data and all auxiliary information to a DAC. The DAC collects data from glider operators (real-time) or from scientists (delayed mode data). In real-time, each DAC converts glider data into a common transport format, typically NetCDF file, according to the glider

data model. It applies the real-time quality controls on the NetCDF files. The DACs push these quality controlled data files to the GDACs. The role of the GDAC is to distribute the best versions of data files.

2.4. ANIMAL-BORNE INSTRUMENTS

In the case of the AniBOS (animal borne ocean sensors) network, the data flow originates with the tag manufacturers, who have well-established infrastructures for decoding, archiving and serving data to their customers.

Data flow and management follow accepted community standards for the data flow from tags to repository (Sequeira *et al.* 2021). PIs who purchase and deploy tags on animals have the option of sharing location and sensor data with the network, and these PIs provide metadata about tag deployments (e.g., location, date, species, tag programming). Additional tag metadata (e.g. sensor sensitivities and calibrations, factory programming defaults, firmware version) are obtained directly from the tag manufacturers.

Real-time data are sent to the regional DACs for standardizing data and metadata into common format and model, and for applying preliminary quality control.
Within 24 hours data is sent to a GDAC, where additional scrutiny can be applied to the data quality control, and data from the different DACs are assembled in one single data repository, forming the real-time AniBOS data product (level 1). A delayed-mode (level 2) data product is then produced by regional data experts and participating PIs.

2.5. FERRYBOXES AND FISHING VESSELS

FerryBox (FB) technology allows taking automated measurements aboard ships of opportunity by means of underway sampling by sensors attached to a sea water intake below the hull. FerryBox systems are installed on commercial vessels by a network of FerryBox contributors, mainly national marine research institutes and environmental agencies.

The national institutes that collect the FerryBox data are in charge of the master data and its versioning. These organisations also have the responsibility to provide their respective regional DACs with access to the open data, but also to communicate larger findings of errors and improvements.

In Europe the FB data flow is organised and coordinated under the EuroGOOS FB Task Team, where EuroGOOS is the European component of the Global Ocean Observing System of the Intergovernmental Oceanographic Commission of UNESCO (IOC GOOS) and the FB Task Team is the expert team running this operational network and promoting scientific synergy and technological collaboration among Europe.

Institutes provides their EuroGOOS regional DACs, i.e. ROOS, with FB data which are shared to European in situ data consumers (CMEMS, EMODnet, etc), together with other in situ data and after undergoing automatic quality control. This happens in real-time (within 24 hours) while delayed mode data are usually made available once the experts complete a second level of quality check.

First steps into a European FerryBox system were taken during an EU-funded project (2002-2005). Since then, a sustainable cooperation between the original and new partners has been established. Currently FerryBox systems are installed on a network of European FerryBox contributors, mainly coastal and marine research institutions and national environmental agencies.

Lately, the Helmholtz-Zentrum Geesthacht (HZG) has been acting as a FB TAC. HZG manages an open and free long-term FB database by using a data model (one dataset for each transect) that can serve operational needs as well as FB community needs. This makes it possible e.g. to compare a set of transects on the same route in an easy way or create so-called scatter plots from one transect over time, or to get a comprehensive graphical overview of sampled FerryBox data in a certain area and at certain routes. Sharing FB with this community may serve to make available advanced FB data products.

Fishing vessels represent an emerging platform network, *i.e.* the so-called Fishery and Oceanography Observing System (FOOS). Data is collected by commercial fishing vessels equipped with an integrated sensor system mounted on fishing gears for collecting data of several parameters, such as position of the fishing operation, depth, water temperature, chlorophyll fluorescence and dissolved oxygen during the haul (to note that number of ships equipped with chemical sensors is yet very limited). Data is usually shared in near real time to the assembly centre that applies data formatting and quality checks before making data available for further uptake and use. Usually, this data is then re-processed for a delayed mode - higher quality and scientific validated version that is included in the reprocessed package (WOD).

2.6 Research Vessels

Data collected during research cruises comprise en-route data acquisition systems, human operations (e.g. physical measurements such as CTD profiles) and the deployment of sensors like ROVs, AUVs or floats.

Cruise data are organised by the PI in charge once the campaign has concluded. PIs are then responsible for transferring data and metadata to data centres.

Because cruise data are managed following a variety of different flows and standards, there could be multiple TACs. For instance, CO_2 and pH data usually follow a route (and feed into the SOCAT initiative), while other biogeochemical data (Chlorophyll, NOx, FOx) may follow other routes (national programmes).

At a national level it is the National Oceanographic Data Centre (NODC) that manages the data collected within research cruises. In Europe, NODCs are networking under the SeaDataNet initiative, which implements the data infrastructure for providing integrated databases of standardised quality from in situ (research cruises) data.

These data are usually delivered in delayed mode and complete the so-called Cruise Summary Reports (CSR) that are the usual means for reporting on cruises or field experiments. Traditionally, it is the Chief Scientist's obligation to submit a CSR to their NODC no later than two weeks after the cruise. This provides a first level inventory of measurements and samples collected at sea.

Two further key partners in the vessels data management are 1) the International Council for the Exploration of the Sea (ICES), which hosts several databases ranging from oceanographic data (including temperature, salinity, oxygen, chlorophyll a, and nutrients measurements) to contaminants, biological effects, and biological community data; and 2) EurOcean (NAUTILOS partner) which is an independent scientific non-governmental organisation aiming at facilitating information exchange and generating value-added products in the field of marine sciences and technologies. EurOcean maintains a Directory of Research Vessels, relying on RV operators providing relevant information.

On a global scale, there are 1) the Partnership for Observation of the Global Oceans (POGO) which is working on improving the sharing of cruises information and related databases and 2) the CLIVAR and Carbon Hydrographic Data Office (CCHDO) that supports access to high quality, global, vessel-based CTD and hydrographic data from GO-SHIP, WOCE, CLIVAR and other repeat hydrography programmes.

2.7 Ocean Acidification

The Surface Ocean CO_2 Atlas (SOCAT) is a synthesis activity for quality-controlled, surface ocean fCO_2 (fugacity of carbon dioxide) observations by the international marine carbon research community. SOCAT enables quantification of the ocean carbon sink and ocean acidification and evaluation of ocean biogeochemical models. SOCAT is hosted by Bjerknes Climate Data Centre (BCDC) and the University of Bergen. The Bjerknes Climate Data Centre has the central role for European and Global data management activities for data related to the Essential Ocean Variable Inorganic Carbon. Currently data management activities for the marine part of the European Research Infrastructure ICOS (Integrated Carbon Observation System) are handled by BCDC. Scientists from the University of Bergen provide data to SOCAT on a global basis.

Some models and recommendations for ocean acidification sensor metadata have been developed for the SOCAT (Pfeil et al., 2013) and the Global Ocean Acidification Observing Network (GOA-ON), and serve as a reference for new initiatives.

Data files available for SOCAT collected by research vessels are first converted to a common file structure. This also includes discarding data not directly relevant for surface ocean CO_2, e.g. meteorological parameters like wind speed and direction, whenever these are supplied in the file. Next, the unit of each parameter is checked and converted into the agreed standard unit, if required. Primary quality control is carried out at this stage. Outliers and unrealistic values in date, time, position, intake temperature, salinity, atmospheric pressure and surface water CO_2 are identified.

Data are then made publicly available, together with information about the investigator, research vessel, Expocode, original cruise naming, metadata (as reported by the investigator), and temporal and geographical coverage.

2.8 European Marine Litter Registry

The first European Marine Litter Database (MLDB) has been developed by EMODnet Chemistry, following the advice of Marine Strategy Framework Directive Technical Group on Marine Litter. The database contains data of beach and sea floor litter from a variety of sources, including existing International and Regional Sea Conventions, and data submitted by EU Member States, EMODnet partners and external research or monitoring projects. A majority of datasets have come from existing

monitoring projects which have published their data in project specific databases (i.e., OSPAR, ICES DATRAS).

The MLDB and floating micro-litter data are accessible via the Data Discovery and Access Service. The Service provides relevant metadata, including identifier, date and location of the survey, data originator and data holding center for all datasets. The data are available for downloading in the EMODnet Chemistry formats, depending on the specific sharing policy applied by the originator.

In addition to the MLDB, EMODnet Chemistry holds data of floating micro-litter from EMODnet partners and external research projects. Floating micro-litter data are gathered in the system using a specific version of ODV format, where relevant characteristics are codified using common vocabularies.

EMODnet Chemistry has elaborated a data flow also for the management of marine litter and plastics data. OGS and the other NODCs collect and format the available marine litter information to populate the EU marine litter database. The formatting is based on fitting marine litter data in the SeaDataNet CDI (Common Data Index) and ODV (Ocean Data View) formats. The SeaDataNet CDI metadata format provides an ISO19115 - ISO19139 based index (metadata base) to individual data sets (such as samples, timeseries, profiles, trajectories, etc), using the SeaDataNet Common Vocabularies and the EDMO (European Directory of Marine Organisations) directory. The CDI format is INSPIRE compliant. The SeaDataNet ODV ASCII data format can be used directly in the ODV, fundamental data analysis and visualisation software.

2.9 Marine Biodiversity and Seabed Habitats

Marine biodiversity data flow starts with in situ measurements, includes archival in data centres and facilitates uptake or redistribution through the relevant European and global databases, such as EMODnet Biology and Seabed Habitats, and OBIS. These systems are well interconnected and the data flow is well established (Figure 2).

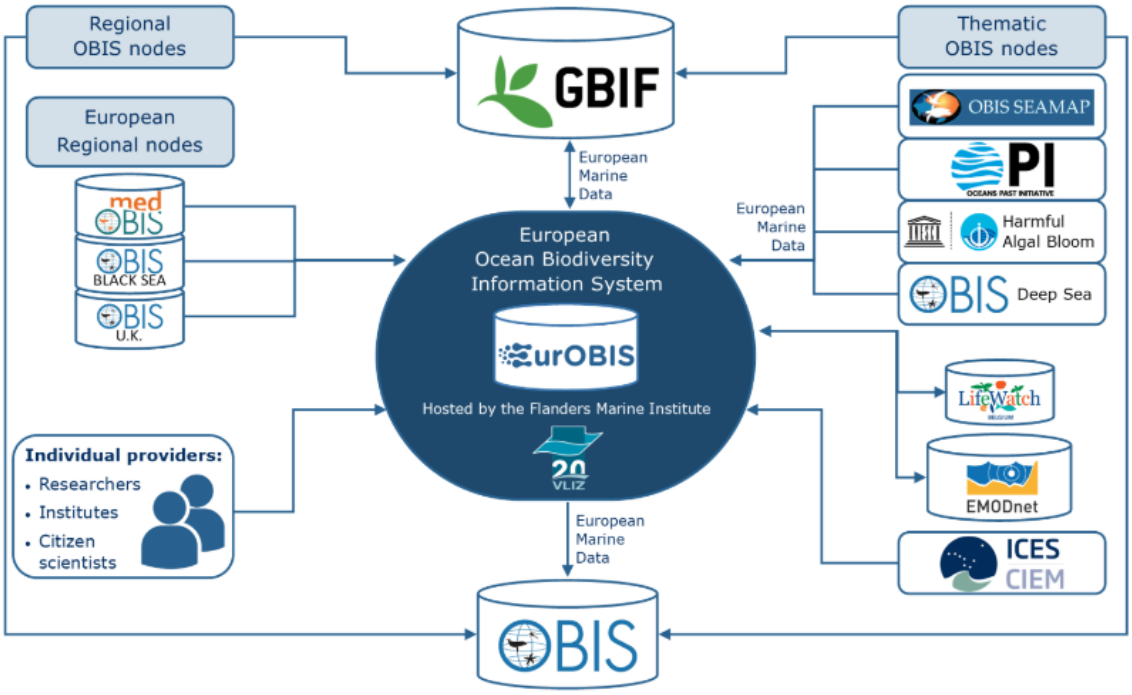

Figure 2 International biodiversity data flow, as adopted by EurOBIS (http://www.eurobis.org/data_flow)

These systems use the Darwin Core standard (https://dwc.tdwg.org/), one of the most commonly used standards for sharing information about biodiversity. Maintained by the organisation Biodiversity Information Standards (formerly TDWG), it provides stable terms and vocabularies for sharing biodiversity data. These include terms related to taxon (e.g. scientificName), identification (e.g. identifiedBy), occurrence (e.g. individualCount), location (e.g. decimalLatitude), event (e.g. habitat), etc.

Biological data can be exchanged using the EMODnet Biology recommended data formats, i.e. as spreadsheets, relational databases or simple text files.

Marine biodiversity information, traditionally acquired by means of visual census methods or samples collection and analysis, is being increasingly supplemented by image-based methods for ecosystem baseline surveys and in repeat monitoring programmes. Image-based surveys are particularly useful to study remote locations, such as the deep sea, where extensive knowledge of the fauna is often limited. Improvements may arise from Open Nomenclature (ON) that is still very little adopted in image-based identifications.

3. NAUTILOS PARAMETERS AND THEMATIC ASSEMBLY CENTERS

Table 1 below summarises NAUTILOS parameters and proposes the main thematic assembly centres and initiatives we recommend NAUTILOS to refer to.

More specifically, the goal for NAUTILOS should be to adopt and adapt some key elements for enabling and facilitating interoperability and NAUTILOS data uptake. These elements provide the reference levels for the data model, the metadata standards and vocabularies and the data transport format.

One goal of NAUTILOS is the development and validation of new sensors or processing techniques which at the end of the workflow have to produce information comparable and aggregable with other sources.

One example is the micro- nano-plastic sensor: plastic litter is a very recent and emerging parameter of interest. Whatever is the technique or methodology to collect the raw data, the final required output should be the litter type, the amount, the location, etc.
This means that NAUTILOS has to identify the best methodology to collect and store raw data for internal processing purposes, and is recommended to deliver data compatible with the litter registry that has been developed under the EMODnet Chemistry project. The same applies to other sensors.

Table 1 NAUTILOS sensors, EOVs and key data integrator infrastructure/initiative.

Sensor(s)	EOV	reference TAC
Deep-ocean CTD	• Salinity • Temperature	hydrography
Silicate electrochemical sensors	Silicate (Inorganic macronutrients)	SeaDataNet EMODnet Chemistry
Phytoplankton and suspended matter sampler	• Phytoplankton biomass and diversity • Stable carbon isotopes • Suspended matter	Ocean Biogeographic Information System (OBIS) and ICES
Dissolved oxygen and fluorescence sensor	• Dissolved oxygen • Ocean colour • Phytoplankton biomass and diversity	Argo Data Management Team
Surface multi-hyperspectral and laser-induced fluorescence imaging sensors	• Ocean colour • Phytoplankton biomass and diversity • Sea surface temperature	To be defined

Ocean acidification sensors	inorganic carbon	EMODnet Chemistry
Submersible nano- and microplastics sampler Low-cost microplastic sensors	Litter, including nano- and microplastics	MSFD Technical Group on Marine Litter & EMODnet Chemistry
Passive broadband acoustic recorder Passive acoustic event recorder	• Ocean sound • Turtles, birds, mammals abundance and distribution • Sea ice	EC Task Group Noise Ocean Biogeographic Information System (OBIS)
Active acoustic profiling sensor	• Suspended particulate organic matter	To be defined
Crowd-sourcing for visual marine image annotations	• Microbe biomass and diversity • Invertebrate abundance and distribution	Ocean Biogeographic Information System (OBIS)
Habitat mapping of key seabed habitats	• Live coral/hard coral cover and composition • Seagrass cover and composition • Sponge habitat cover	Ocean Biogeographic Information System (OBIS) EMODnet Seabed Habitats

One parameter that may require a special focus and further analysis is the ocean sound. NAUTILOS is developing an innovative multi-purpose sensor to match both acoustical oceanography (particle detection, anthropogenic noise detection) and bioacoustics (dolphin and porpoise clicks). While in the case of particle detection or bioacoustics the final outcome could refer to an already existing best practice, in case of anthropogenic noise detection it is recommended to save raw data enriched with medata with full information about the sensor, the sampling methodology, the low level processing methods etc, to enable an *a posteriori* reanalysis and uptake of data. This matches well with the latest recommendations from the scientific community.

For instance, Martinez *et al.* (2021) proposed a universal architecture for in situ, real-time ocean sound monitoring, compliant with the needs of the ocean observing community (i.e. MSFD indicators) and following best practices on underwater sound measurement methodologies. They propose a metadata-driven approach in which metadata are not added to the acquired data once gathered, but prepared beforehand and control the acquisition process. Using this approach, metadata may not only reflect what has been measured, but also unambiguously define the whole acquisition chain, including

sensor setup, signal processing, formatting, etc. Thus, the acquisition chain in a deployment can be easily replicated based on its metadata.

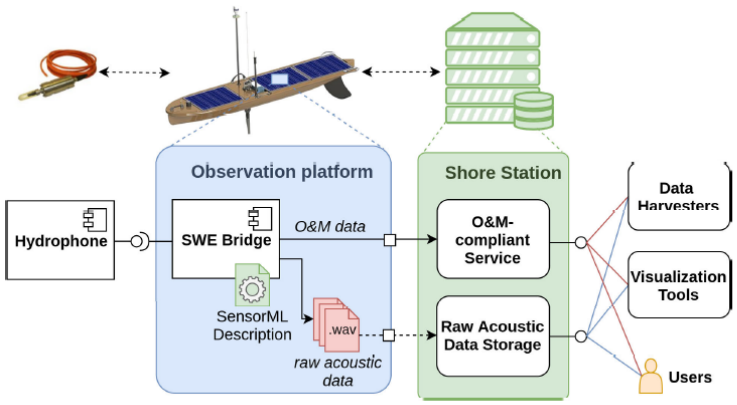

Figure 3 Proposed architecture for ocean sound monitoring (Martinez et al., 2021)

Then, the streams of data are sent to the shore station in real-time or delayed (depending on the telemetry used) to be published by means of a generic (big) data storage service such as File Transfer Protocol (FTP) or ERDDAP.

4. CONCLUSIONS AND RECOMMENDATIONS

The technology is evolving in multiple contexts ranging from miniaturization of sensors and the evolution of power-harvesting systems for platforms, to improve data-transmission systems etc. Concurrently, new technical approaches to data analysis are emerging, including cloud computing, big data analytics, machine learning and AI.

NAUTILOS project is largely contributing to this process, anyhow the harmonization of technologies, methodologies and procedures is a vital step in ensuring efficiency and optimal returns from any kind of instrument, employed on a transnational level.

This is because such harmonization leads to an efficient use of resources and information, improves the consistency of services and products, and helps to provide uniformed protocols, thereby allowing intercomparisons of, and conclusions from different sensors.

To this end, an analysis of data management practices of different sensors and systems has been performed. Key elements of thematic data assembly and processing centers have been identified and NAUTILOS back end infrastructure is compatible and in line with international recommendations to implement interoperability and FAIR principles (see also NAUTILOS Data Management Plan and Deliverable 8.3).

Despite being generated by new sensors, the data produced within NAUTILOS will be used by well-defined programmes and applications. Therefore, NAUTILOS data need to be managed by applying common standards (in particular common vocabularies and common data transport format). This is the cornerstone of interchange with the end users.

Nevertheless, for those innovative parameters lacking standardised data processing (e.g. acoustic profiling sensors), we suggest creating higher resolution datasets enriched with full set of metadata including information about the sensor, the low level processing methodology, file format description, etc. that would facilitate *a posteriori* reanalyses and user uptake.

In the case of image annotations, for instance, a first level of dataset would contain information about location and species taxonomy, and a second one would include the data item itself, i.e. the video or image sequence. This would greatly benefit any future reanalysis of the data acquisition.

Furthermore, managing all data flows per individual ‘mission’ would enable assigning a doi. An efficient way to achieve this would be to assign a doi to the parameter-platform-mission match.

The next steps will focus on the definition of homogeneous best practices on the different aspects described above. The final objective is now to reach some consensus on methods and best practices in the utilization and deployment of the sensors.

APPENDIX 1: REFERENCES AND RELATED DOCUMENTS

Reference	Source or Link/Location
AniBOS Proposal to form the GOOS network, Animal Borne Ocean Sensors (2020)	
Bourtzis, Tilemachos. (2015). The Role of Marine Data in Advancing Development.	https://www.researchgate.net/publication/283455113_The_Role_of_Marine_Data_in_Advancing_Development
Buga, G. Sarbu, L. Fryberg, K. Wesslander, J. Gatti, S. Iona, M. Tsompanou, M. M. Larsen, A.K. Østrem, M. Lipizer, M.E. Molina Jack, A. Giorgetti, 2021 Quality Control steps for EMODnet Chemistry Eutrophication aggregated datasets - v2021, 10/03/2021	https://www.emodnet-chemistry.eu/doi/documents/Eutrophication_QC_Steps_Collection2021.pdf
EGO gliders data management team (2021). EGO gliders NetCDF format reference manual.	https://doi.org/10.13155/34980
Giorgetti A, Lipizer M, Molina Jack ME, Holdsworth N, Jensen HM, Buga L, Sarbu G, Iona A, Gatti J, Larsen M, Fyrberg L, Østrem AK and Schlitzer R (2020) Aggregated and Validated Datasets for the European Seas: The Contribution of EMODnet Chemistry. Front. Mar. Sci. 7:583657.	10.3389/fmars.2020.583657
Horton T, Marsh L, Bett BJ, Gates AR, Jones DOB, Benoist NMA, Pfeifer S, Simon-Lledó E, Durden JM, Vandepitte L and Appeltans W (2021) Recommendations for the Standardisation of Open Taxonomic Nomenclature for Image-Based Identifications. Front. Mar. Sci. 8:620702.	doi: 10.3389/fmars.2021.620702
Ignaszewski, M. (2018). Description of the Argo GDAC File Checks: Data Format and Consistency Checks (Brest: Ifremer). doi: 10.13155/46120	https://doi.org/10.13155/46120
J. del Rio, D. M. Toma, T. C. O'Reilly, A. Broring, D. R. Dana, F. Bache, K. L. Headley, A. Manuel-Lazaro, and D. R. Edgington, ''Standards-based plug & work for instruments in ocean observing systems,'' IEEE J. Ocean. Eng., vol. 39, no. 3, pp. 430–443, Jul. 2014.	
Lindstrom, E., Gunn, J., Fischer, A., McCurdy, A., and Glover, L. K. (2012). A Framework for Ocean Observing.	http://www.oceanobs09.net/foo/FOO_Report.pdf
Martínez, A. García-Benadí, D. M. Toma, E. Delory, S. Gomáriz and J. Del-Río, "Metadata-Driven Universal Real-Time Ocean Sound Measurement Architecture," in IEEE Access, vol. 9, pp. 28282-28301, 2021	doi: 10.1109/ACCESS.2021.3058744.
Pfeil, Benjamin, et al. "A uniform, quality controlled Surface Ocean CO_2 Atlas (SOCAT)." Earth System Science Data 5.1 (2013): 125-143.	doi:10.5194/essd-5-125-2013
Roemmich, D., and J. Gilson. 2009. The 2004–2007 mean and annual cycle of temperature, salinity and steric height in the global ocean from the Argo Program. Progress in Oceanography 82(2): 81–100.	

Sastri AR, Christian JR, Achterberg EP, Atamanchuk D, Buck JJH, Bresnahan P, Duke PJ, Evans W, Gonski SF, Johnson B, Juniper SK, Mihaly S, Miller LA, Morley M, Murphy D, Nakaoka S-i, Ono T, Parker G, Simpson K and Tsunoda T (2019) Perspectives on in situ Sensors for Ocean Acidification Research. Front. Mar. Sci. 6:653.	doi: 10.3389/fmars.2019.00653
Sequeira, AMM, O'Toole, M, Keates, TR, et al. A standardisation framework for bio-logging data to advance ecological research and conservation. Methods Ecol Evol. 2021; 12: 996– 1007.	https://doi.org/10.1111/2041-210X.13593
Tanhua, T., McCurdy, A., Fischer, A., Appeltans, W., Bax, N., Currie, K., ... & Wilkin, J. (2019). What we have learned from the framework for ocean observing: Evolution of the global ocean observing system. Frontiers in Marine Science, 6, 471.	https://doi.org/10.3389/fmars.2019.00471
Wong, A. P., et al. "Argo data 1999–2019: two million temperature-salinity profiles and subsurface velocity observations from a global array of profiling floats." (2020).	https://doi.org/10.3389/fmars.2020.00700
Wong, A., Keeley, R., and Carval, T. Argo Data Management Team (2020). Argo Quality Control Manual for CTD and Trajectory Data, V3.3 (Brest: Ifremer). doi: 10.13155/33951	https://archimer.ifremer.fr/doc/00228/33951/