



Marine Microbial Biodiversity, Bioinformatics & Biotechnology



Grant agreement n°287589

Acronym : Micro B3

Start date of project: 01/01/2012, funded for 48 month

Deliverable 3.1

Review of oceanographic data services for Micro B3

Version: 1.1

Circulated to: Extended Executive Board (September 2012)

Approved by:

Expected Submission Date: 30.06.2012

Actual submission Date: 03.10.2012

Lead Party for Deliverable: MARIS

Mail: dick@maris.nl

Tel.: +31 70 3004710

Dissemination level:

Public (PU)	
Restricted to other programme participants (including the Commission Services) (PP)	
Restricted to a group specified by the consortium (including the Commission Services) (RE)	X
Confidential, only for members of the consortium (including the Commission Services) (CO)	



The Micro B3 project is funded from the European Union's Seventh Framework Programme (Joint Call OCEAN.2011-2: Marine microbial diversity – new insights into marine ecosystems functioning and its biotechnological potential) under the grant agreement no 287589. The Micro B3 project is solely responsible for this publication. It does not represent the opinion of the EU. The EU is not responsible for any use that might be made of data appearing herein.





Summary

The Micro B3 project aims for a better understanding of the complexity of marine microbial communities and their role in climate change. This requires that the data sets and information on marine organisms and genes are complemented with their environmental context. Oceanographic and marine environmental data will be provided to Micro B3 by major existing oceanographic data management infrastructures, such as SeaDataNet, EMODNet, ICES, EurOBIS and PANGAEA from existing ocean and marine data collection activities from multiple sources. This will be done as part of WP3. Moreover data will be collected in the framework of Micro B3 via the **Ocean Sampling Day (OSD)** (WP2) and derived from the **Tara Oceans** expedition for which WP3 will give data management support. These campaigns not only provide environmental data but also genomic samples. WP4 is charged with drafting an **Ocean Sampling Handbook**. This requires input from both the genomic community and the oceanographic community. Finally WP5 is charged with building the Micro B3 Information System (MB3-IS) to provide the bioinformatics capacity for marine biodiversity data processing, analysis and biotechnological exploitation. This requires data input from both the genomic data infrastructure (EMBL-EBI) and the ocean environmental data infrastructure as presented by SeaDataNet, ICES, EurOBIS and PANGAEA.

Fulfilling the oceanographic data needs of Micro B3 requires an analysis of the demand, a confrontation with the existing oceanographic data provision, and a functional analysis of the way that the flow of data, both ocean environmental and genomic data, from the field via the data management infrastructures to Micro B3, MB3-IS and users might be structured and organised.

This analysis is undertaken by WP3 in interaction and cooperation with WP4 and WP5 in particular and is ongoing because it also depends on the progress of these WPs. Therefore this Deliverable 3.1 gives a review of the oceanographic data services for MicroB3, while the further analysis will be reported in Deliverable 3.3 later this year.



Table of Contents

1.0 Context and objectives of WP3 activities

2.0 European oceanographic data management infrastructure

- 2.1 SeaDataNet**
- 2.2 EMODNet**
- 2.3 EurOBIS and WoRMS**
- 2.4 ICES**
- 2.5 PANGAEA**

1.0 Context and objectives of WP3 activities

The Micro B3 project aims for a better understanding of the complexity of marine microbial communities and their role in climate change. This requires that the data sets and information on marine organisms and genes are complemented with their environmental context.

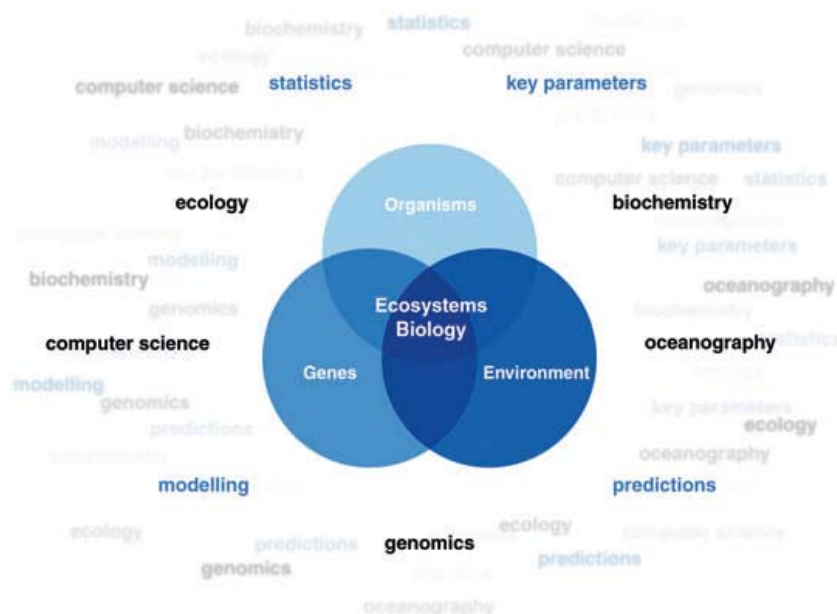


Figure: Integrating the information on the diversity of the organisms with their potential functions, reflected in their genes, and the environmental conditions surrounding them.

Oceanographic and marine environmental data include a very wide range of measurements and variables covering a broad, multidisciplinary spectrum of projects and programmes. Oceanographic and marine data are collected by a multitude of research institutes, governmental organizations and private companies. Various heterogeneous observing sensors are used installed on research vessels, submarines, aircraft, moorings, drifting buoys and satellites. These sensors measure physical, chemical, biological, geological and geophysical parameters, with further data resulting from the analysis of water and sediment samples for a wide variety of parameters.

Data on marine organisms and samples for genomic analyses are taken by marine research and ecological stations such as coastal and offshore laboratories and by research vessels during scientific cruises.



These existing and ongoing data flows and their long term archives provide a great potential input for the Micro B3 community and its users.

In addition, Micro B3 as part of WP2 is organising **Ocean Sampling Day (OSD)** which will be a special campaign with multiple sites in the global ocean to collect samples at the same day which will be placed on the summer solstice (June 21st) in the year 2014. Already more than 30 research groups have expressed interest in providing samples or to give access to sampling sites.

OSD will focus on collecting samples for genomic analyses as well as environmental data. To ensure maximum usefulness of these samples it is planned to do the OSD campaign across all sites using agreed upon best practices developed within Micro B3. Therefore as part of WP4 an Ocean Sampling Handbook is being drafted. In particular, all sites will be expected to confirm to the minimum information checklists of the Genomic Standards Consortium for describing metagenomic samples, such as geo-reference, date and time of sampling. Moreover the checklists include environmental parameters, like temperature, depth, pH etc. thereby following best practice in oceanographic data collection.

The handbook will then guide sampling and analysis groups on best practice for Micro B3. The handbook is being produced with the Ocean Sampling Day as principal target, but it will have a wider utility beyond Micro B3 and therefore it will be handed over to an appropriate body for wider dissemination and long-term maintenance, such as the Genomics Standards Consortium and possibly also the International Oceanographic Data and Information Exchange programme (IODE) of the Intergovernmental Oceanographic Commission (IOC) of UNESCO that maintains and distributes standards in the oceanographic environmental data domain.

This illustrates one of the major challenges in Micro B3 which is to establish cooperation and interoperability between the oceanographic research community and the genomics research community. Both communities have their own practices and infrastructures for acquiring, processing and managing data, but Micro B3 aims to arrange that users within Micro B3 and external users of Micro B3 will have efficient access to both the genomic data and the associated environmental data. For the latter it is important to have access to environmental data at the location and time of the genomic sample taking, but also to longer term time series or climatologies of specific environmental parameters in the area of the location, both from in-situ and satellite observations.

As part of WP5 Micro B3 will build the Micro B3 Information System (MB3-IS) to provide the bioinformatics capacity for marine biodiversity data processing, analysis and biotechnological exploitation. It will support users in the complex process of handling the



sequence data arising from studies of a range of organisms (protists, viruses and prokaryotes), both in isolation and in mixed organism samples, and a variety of sequencing platforms whilst integrating interpretations with related environmental data.

The MB3-IS will support five generalized use cases and it will have a data integration component to retrieve contextual environmental (meta)data obtained from in-situ and remote sensing measurements and to integrate these with the quality controlled and processed genomic data. For this purpose, this component should have well established exchange mechanisms with existing data infrastructures both in the genomics domain and in the oceanographic environment domain.

Furthermore there is a European sampling campaign ongoing, **Tara Oceans**, that will contribute to the data provision for Micro B3. This expedition aims to study for the first time on a global scale, the effects of climate change on marine microorganisms, such as plankton, from which originated all living organisms on our planet. Marine organisms make up approximately 90% of the total ocean biomass, absorb the majority of the atmospheric carbon dioxide, and produce half of our planet's oxygen. Modern techniques and methods are also used to evaluate the biodiversity and activity in the planktonic samples, which are collected in different types of ecosystems in the oceans. The adaptation of the plankton to a rapidly changing earth system will also be assessed.

All the data generated throughout the **Tara Oceans** project will form an open-source multidimensional bio-oceanographic database that will allow generating predictive models of the spatio-temporal evolution of plankton ecosystems. For the latter Micro B3 will provide data management support to Tara Oceans for ensuring that all collected data and samples are well stored and documented in an appropriate data management system following best practices in the oceanographic and the genomics domains.

This Deliverable is drafted as part of WP3. This Micro B3 Work Package has a primary focus on arranging and establishing a well structured and operational provision of oceanographic environmental data to the Micro B3 project. Moreover WP3 is aiming at contributing and transferring best practices from the oceanographic environmental domain to other relevant WPs in Micro B3:

- WP4 for integrating oceanographic data acquisition and management standards and procedures in the development of the Ocean Sampling Handbook
- WP4 for arranging the flow of data from field acquisition to the data management infrastructures respectively for genomics, operated by the European Molecular Biology Laboratory - European Bioinformatics Institute (EMBL - EBI) and the



European ocean data management infrastructure and for establishing interoperability solutions between these large infrastructures on behalf of the user community in Micro B3 and beyond

- WP2 for ensuring that the oceanographic environmental data sets as collected during OSD get populated in the data management infrastructures for wider use; also in a comparable way for the data as collected during the Tara Oceans expedition
- WP5 for arranging that existing oceanographic environmental data can be retrieved and made available for MB3-IS by an efficient exchange mechanism from the European ocean data management infrastructure.

Fulfilling the oceanographic data needs of Micro B3 thus requires an analysis of the demand, a confrontation with the existing oceanographic data provision, and a functional analysis of the way that the flow of data, both ocean environmental and genomic data, from the field via the data management infrastructures to Micro B3, MB3-IS and users might be structured and organised. This analysis is undertaken by WP3 in interaction and cooperation with WP4 and WP5 in particular and is ongoing because it also depends on the progress of these WPs.

This Deliverable D3.1 is drafted as part of Task 3-1: Establishing interoperability between Micro B3 and the Oceanographic Environmental data management systems. As a first step it gives a review of the oceanographic data services for MicroB3, while the further analysis and results of tuning with related WPs will be reported in Deliverable 3.3 later this year, also incorporating then a technical analysis and specifications of the interoperability solutions. The actual implementation of the interoperability solutions and the operational provision of oceanographic data to Micro B3 will take place in the second and third year of the Micro B3 project.

2.0 European oceanographic data management infrastructure

In Micro B3 there is a need to store and to have access to the oceanographic environmental parameters that are collected during Ocean Sampling Day (OSD) and during the Tara Oceans expedition. Moreover there is a need for long term in-situ time series or climatologies and remote sensing data sets in the area of the OSD and Tara Ocean sampling locations. These have to be retrieved from existing oceanographic monitoring and scientific observation activities.

Oceanographic and marine environmental data include a very wide range of measurements and variables covering a broad, multidisciplinary spectrum of projects and programmes. Oceanographic and marine data are collected by a multitude of research institutes, governmental organizations and private companies. Various heterogeneous observing sensors are used installed on research vessels, submarines, aircraft, moorings, drifting buoys and satellites. These sensors measure physical, chemical, biological, geological and geophysical parameters, with further data resulting from the analysis of water and sediment samples for a wide variety of parameters.

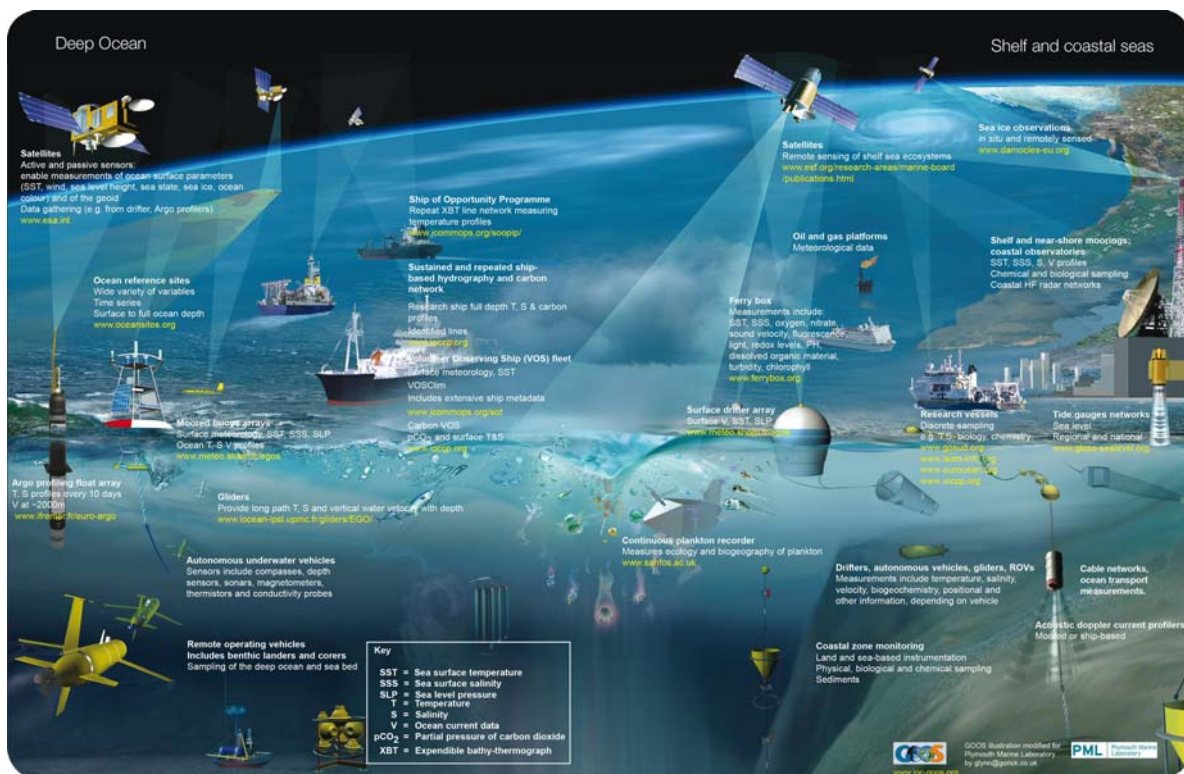


Figure: overview of oceanographic observations - derived from EMODNet Vision Paper of ESF Marine Board

A recent EMODNet study estimates that Europe spends annually 1.4 Billion Euro on data acquisition by in-situ (1.0 B) and satellite platforms (0.4 B)

There are a number of major initiatives and special data centres in Europe for managing and giving access to the large existing collections of environmental data:

- SeaDataNet, pan-European infrastructure for marine environmental data
- EMODNet, European Marine Observation and Data Network
- EurOBIS, European Ocean Biogeographic Information System
- ICES, International Council for the Exploration of the Sea
- PANGAEA, Data Publisher for Earth & Environmental Science

It is planned that these infrastructures will work together to provide a large range of data sets that will be fit for serving the needs of Micro B3. Therefore as a first step these infrastructures will be reviewed.

2.1 SeaDataNet

SeaDataNet (<http://www.seadatanet.org>) is the leading network in Europe, actively operating and further developing a Pan-European infrastructure for managing, indexing and providing access to ocean and marine data sets and data products, acquired from research cruises and other observational activities in European marine waters and global oceans. It connects the National Oceanographic Data Centres (NODCs), and marine information services of major research institutes, from 35 coastal states bordering the European seas, and also includes IOC-IODE, ICES and EU-JRC in its network. These data centres work together on refining their standards and expanding their infrastructure and associated services. These activities are continued with EU support in the FP7 SeaDataNet II project (2011 - 2015), succeeding the earlier FP5 Sea-Search project (2002 - 2005) and FP6 SeaDataNet project (2006 - 2011).

The SeaDataNet data centres are highly skilled and have been actively engaged in marine data management for many decades and have the essential capabilities and facilities for data quality control, long term stewardship, retrieval and distribution of marine and ocean data. They are departments of the key marine research institutes and marine management organisations in Europe. Moreover, they maintain national networks, communicating with other marine and ocean institutes in their country, and promoting SeaDataNet standards for QA/QC, storage, catalogues, and access. Together they are very active from the shore to the deep ocean, for marine research and environmental monitoring, including research activities in different themes (e.g. climate change, marine hydrology and chemistry, hydrodynamics, geology, marine living resources, biodiversity and habitats).

SeaDataNet has focused, with success, on establishing common standards and on applying those standards for interconnecting the data centres enabling the provision of integrated online access to comprehensive sets of multi-disciplinary, *in situ* and remote sensing marine data, metadata and products. The SeaDataNet architecture has been designed as a multidisciplinary system from the beginning. It is able to support a wide variety of data types and to serve several sector communities. SeaDataNet is also actively sharing its technologies and expertise, spreading and expanding its approach, and building bridges to other communities in the marine domain. This has resulted in adoption and an active role for a number of SeaDataNet partners in related data management projects, for example, the FP7 projects Geo-Seas, Upgrade Black Sea SCENE, EUROLLEETS, and JERICO, as well as close cooperation with EuroGOOS and MyOcean (GMES):

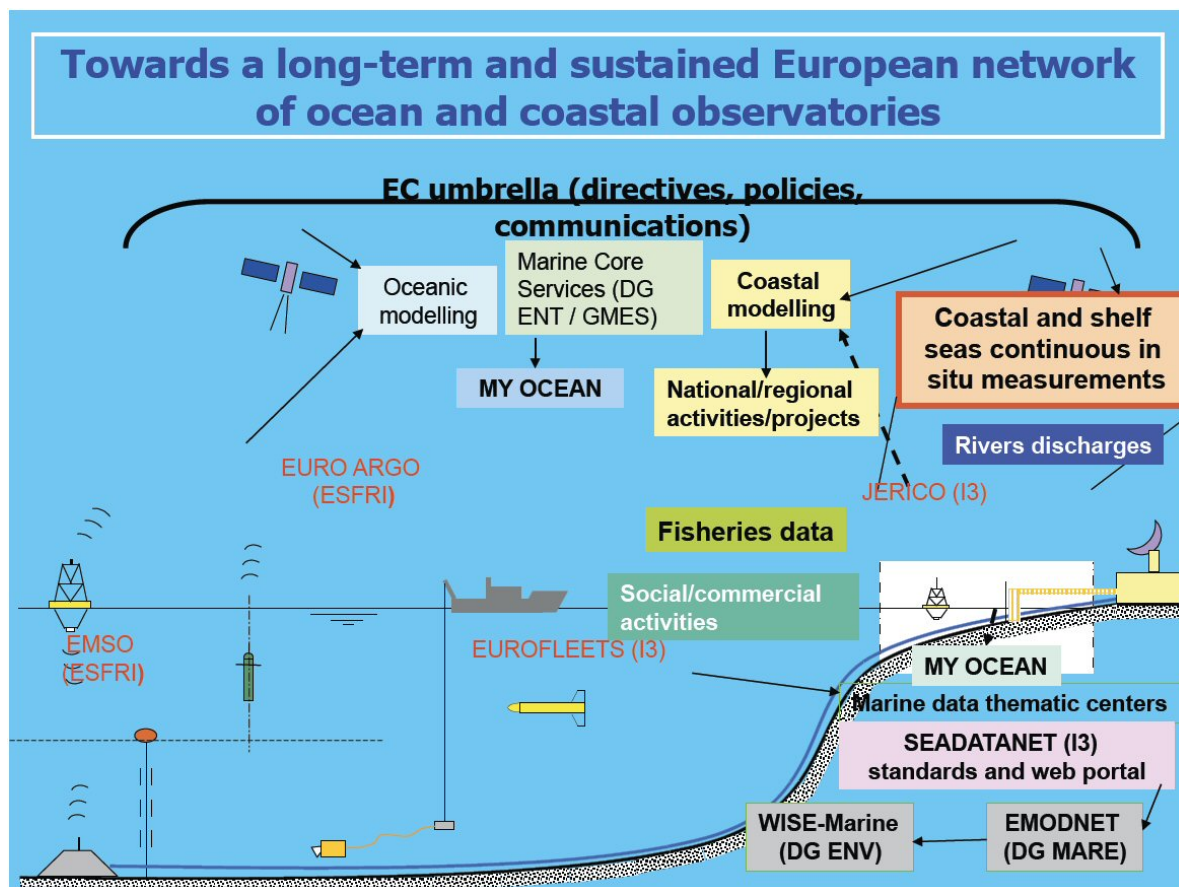


Figure: overview of European initiatives for ocean observations and interaction with SeaDataNet and EMODNet for data management

SeaDataNet provides online unified access to distributed datasets *via* its portal website to the vast resources of marine and ocean datasets, managed by the distributed data centres. The **Common Data Index (CDI)** is the key Discovery and Delivery service. It gives users a highly detailed insight in the geographical coverage, and other metadata features of data across the different data centres, based upon the ISO 19115 content model. Users can request access to identified datasets in a harmonised way, using a shopping basket. They can follow the processing of requests *via* an online transaction register and can download

datasets in the SeaDataNet standard formats. At present the CDI service provides metadata and access to more than 1 million data sets, originating from more than 375 organisations in Europe, and more than 65 connected data centres, covering physical, geological, chemical, biological and geophysical data, and acquired in European waters and global oceans.

Access is given as downloading services, whereby the data sets can be downloaded by users from the data centres in the SeaDataNet standard data exchange formats. When a user wants access to the actual data sets via the shopping cart mechanism, then a user has to login with his/her SeaDataNet user id - password. This facilitates the further processing of a shopping request for multiple datasets to multiple data providers by one user action. Each dataset has a data access restriction determined by its data provider and chosen from a common vocabulary of options varying from unrestricted to restricted with a number of options in between. This restriction combined with the registered role of the SeaDataNet user determines whether the user gets immediate access to the requested dataset (within circa 10 - 15 minutes) or has to discuss terms with the data provider (within 1 - 3 days) or is denied access. Once granted access the user can download the agreed datasets using an online Transaction Register and its SeaDataNet user id - password. The user registration is thus required for facilitating the differentiated access policies of data providers, for enabling asynchronous processing of multiple requests to multiple data providers, and for providing data providers administrative information about their users which is used to improve services in contact with users and to justify their services to government and public.

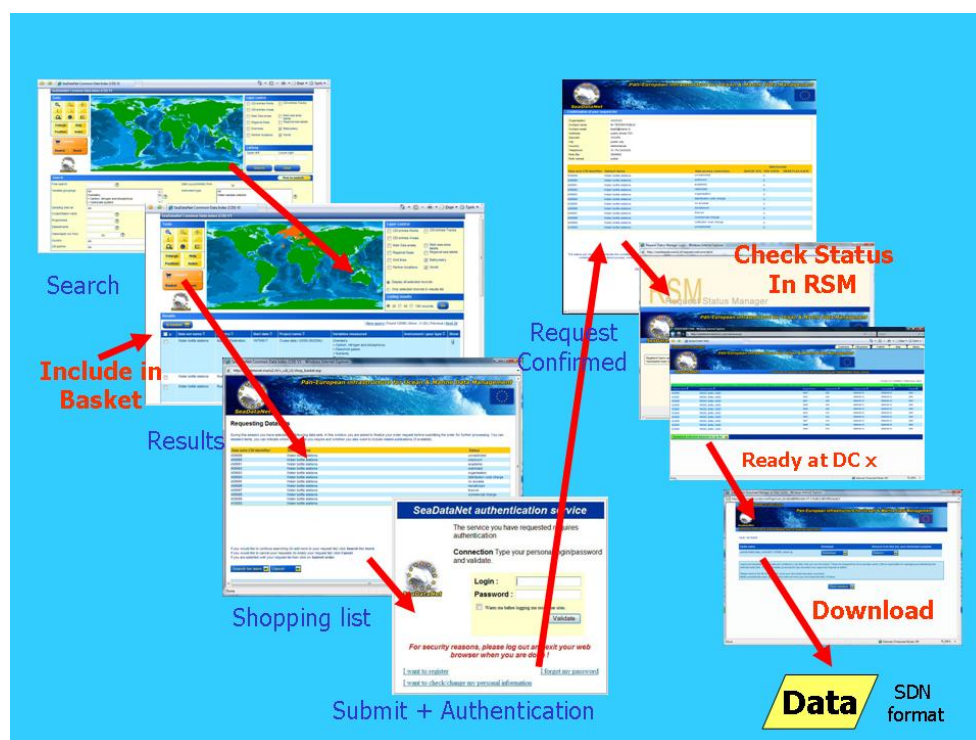


Figure: CDI data discovery and access service - shopping dialogue

The CDI Data Discovery and Access Service user interface includes a mapping service, that supports OGC-WMS for adding external map layers and exchanging the CDI map layers with external systems.

It is encouraged that all data becomes available with the SeaDataNet license, which in practice means that registered users have immediate access. But the data access restriction label enables data providers to apply more restrictive conditions where they feel appropriate. This is used for instance to main a moratorium period for scientific data; it is also used by a number of environmental monitoring agencies to regulate access to recent environmental data.

The SeaDataNet infrastructure comprises thus a network of interconnected data centres and a central SeaDataNet portal. The portal provides users access to the range of metadata directories, the CDI data discovery and access service, a range of data products and the various SeaDataNet standards and tools.

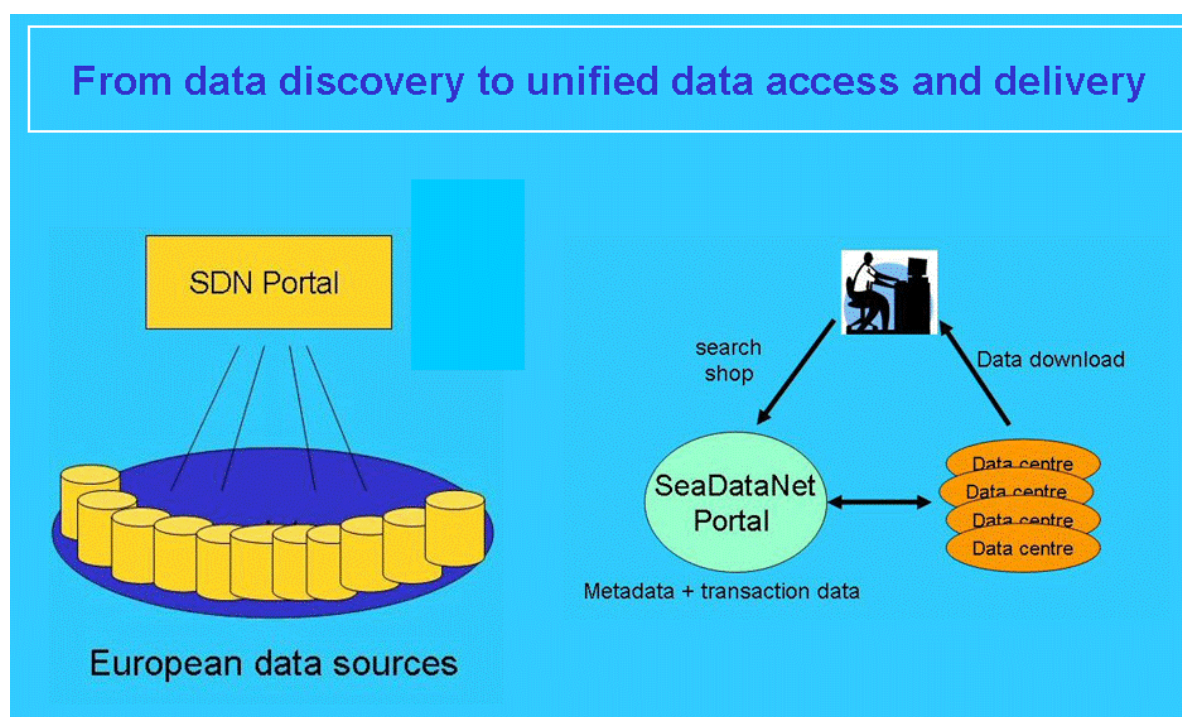


Figure: Central portal with connected data centres and search-find-bind process

The SeaDataNet metadata directories provide overviews of marine organisations in Europe and their engagement in research, scientific cruises, monitoring and data management for European waters and global oceans:

- EDMO: European Directory of Marine Organisations (>2200 entries)
- EDMED: European Directory of Marine Environmental Data sets (>3000 entries)
- EDMERP: European Directory of Marine Environmental Research Projects (>2500 entries)



- CSR: Cruise Summary Reports (>31500 entries)
- EDIOS: European Directory of Ocean-observing Systems (>270 programmes for the UK alone and many underway for other European countries)

These Directories are maintained by NODCs for their country and published at pan-European level

SeaDataNet maintains and operates Common Vocabulary Web services, covering a broad spectrum of ocean and marine disciplines. The common terms are used to mark up metadata, data and data products in a consistent and coherent way. Governance is regulated by an international board. At present the Vocabulary Services comprise over 120000 terms in over 100 lists.

SeaDataNet provides documentation and common software tools for metadata and data formatting, Quality Control - Quality Assurance, statistical analysis (DIVA) and a versatile software package for data analysis and presentation (ODV). These tools can be downloaded without any restriction from the SeaDataNet portal.

SeaDataNet develops and publishes standard data products for maritime regions Arctic waters, North Sea, North Atlantic Ocean, Baltic Sea, Mediterranean Sea, and Black Sea. These include gridded fields for public use and large aggregated data sets for specific user communities, for selected data types. Gridded fields (such as climatologies) are relevant for applications which cannot directly make use of observation data that are generally sparse and heterogeneously distributed. These applications are numerous, including initialisation, calibration and validation of ocean models (in support of projects like MyOcean), analyses of changes and trends at seasonal, annual and interannual time scales and budget analyses (such as heat content and total biomass), and preparatory phases of environmental assessment (MSFD). Large aggregated data sets for selected parameters are relevant input for modelling, such as applied within MyOcean. These products are prepared using SeaDataNet data sets and the DIVA and ODV software tools. Online publishing is done via the OceanBrowser application at the SeaDataNet portal. These products can be downloaded as NetCDF files and can be shared as map layers by means of OGC-WMS.

Architecture

The SeaDataNet infrastructure comprises the following middleware services:

- Discovery services = Metadata directories and User interfaces
- Vocabulary services = Common vocabularies and Governance
- Security services = Authentication, Authorisation & Accounting
- Delivery services = Requesting and Downloading of data sets
- Viewing services = Mapping of metadata



- Product viewing services = Viewing of generic and standard products in maps
- Monitoring services = Statistics on system usage and performance and Registration of data requests and transactions
- Maintenance services = Entry and updating of metadata by data centres

In SeaDataNet II the infrastructure is being extended with a further development of:

- Viewing services = Quick views and Visualisation of data
- Machine to Machine interface services = Providing interoperability and automatic exchanges with other systems
- Delivery services = Requesting and downloading of aggregated data sets
- Delivery services = Access to real-time oceanography data

SeaDataNet has developed and applies standards such as :

- Common metadata standards and XML schemas, based on ISO 19115
- Standard data transport formats ODV ASCII, MEDATLAS and NetCDF (CF)
- Common quality control methods and quality flag scale
- Common Vocabulary Web services, used to mark up metadata and data, covering a broad spectrum of disciplines. Governed by an international board (SeaVox)
- SOAP Web services for various communication tasks
- OGC services (WMS, WFS) for viewing services of map layers and data products

In SeaDataNet II further activities are undertaken for achieving full INSPIRE compliance, including a further refinement of the SeaDataNet Common Data Index (CDI) metadata profile to ISO 19139. And establishing common SeaDataNet profiles of OGC standards such as SensorML, Observations & Measurements (O&M) schema and Sensor Observation Service (SOS) for supporting real-time oceanographic data provision and streamlining the flow and standard description of signals from the *in situ* instruments (e.g. buoys, floats, drifters, gauges, gliders, ferrybox) to the data management system. Furthermore it is explored whether the Authorization Authentication and Accounting (AAA) system for controlled access to data can be expanded with OpenID support.

In addition work is ongoing for developing machine-to-machine interfacing as part of the CDI Data Discovery and Access Service service next to the present user interfaces. In particular it will be made possible to formulate a query by OpenSearch protocol (attribute, spatial and temporal selection) on another system that then will be processed via machine-to-machine interfacing by the CDI service. This process will be supported by SOAP webservice. This extra interfacing is aiming at providing regular data services for a number of established communities such as MyOcean and WISE-MARINE (EEA and EU DG



Environment). Micro B3 in effect requires a somewhat comparable data service that might be arranged following the same technical principles.

SeaDataNet plays a leading role in the further development and actual implementation of the overarching European Marine Observation and Data Network (EMODNet) that is formulated in the Marine Strategy Framework Directive (MSFD).

2.2 EMODNet

Implementation of the **Marine Strategy Framework Directive (MSFD)** will be aided by an overarching **European Marine Observation and Data Network (EMODNet)**. This will be a network of existing and developing European observation systems, linked by a data management structure covering all European coastal waters, shelf seas and surrounding ocean basins. It must facilitate long-term and sustainable access to the high-quality data necessary to understand the biological, chemical and physical behaviour of seas and oceans.

EMODNet will underpin, and provide data to **WISE-Marine**, the marine component of the EEA's Shared Environmental Information System (SEIS). WISE-Marine is intended to fulfil the reporting obligations of the Marine Strategy Framework Directive and to inform the European public on indicators for Good Environmental Status of sea basins. EMODNet is coordinated at EU level with the other European directives (**INSPIRE**) and large-scale framework programmes on European and global scales (**GMES and GEOSS**), that urge access to, and exchange of, environmental data and information.

The SeaDataNet infrastructure and standards have been adopted as core for the EMODNet data management component. Partnerships from the SeaDataNet consortium successfully bid to develop a number of the EMODNet preparatory actions. These are undertaken to test the "proof of concept" of EMODnet. Portals for a number of maritime basins have been set up for hydrographic, geological, biological, chemical and physical data as well as functional habitat maps. These portals provide access to marine data and data products of a standard format and known quality. In practice most of the portals (chemistry, hydrography, physics, geology (via link with Geo-Seas)) have adopted the SeaDataNet approach of using the CDI data discovery and access service including its flexible data access restrictions for giving overview and access to basic measurements datasets.

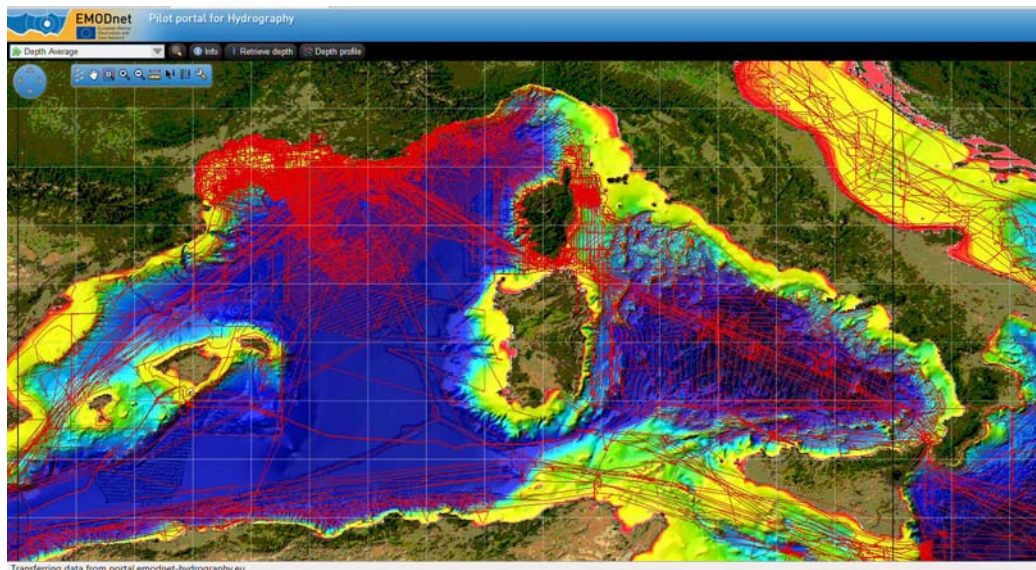


Figure: High Resolution Digital Bathymetry as part of EMODNet Hydrography

EMODNet encourages more data providers to come forward for data sharing and participating in the process of making complete overviews and homogeneous data products. This will give wider visibility at the policy and management levels both at EU and Member States that should seek integration of EMODnet output and services in management and policy processes and that will decide upon its future sustained funding.

In parallel further RTD work as taking place in SeaDataNet II will and must continue on standards and protocols that can be applied as basis for the expanding EMODNet portals. For this purpose Horizon 2020 will provide a fertile ground for SeaDataNet III and Geo-Seas II projects that will refine and expand the data management standards as well as the services.

2.3 EurOBIS and WoRMS

The European Ocean Biogeographic Information System—EurOBIS—is an integrated data system for the EU Network of Excellence “Marine Biodiversity and Ecosystem Functioning” (MarBEF) in 2004. Its principle aims are to centralise the largely scattered biogeographical data on marine species collected by European institutions and to make these quality-controlled data freely available and easily accessible. It is a distributed system in which individual datasets go through a series of quality control procedures. EurOBIS is freely available online at www.eurobis.org, where marine biogeographical data—with a focus on taxonomy, temporal and spatial distribution—can be consulted and downloaded for analyses. It is operated and managed by VLIZ - Belgium.

The database consists of a standard list of data fields, the OBIS schema version 1.1, which is an extension of DarwinCore 2. This OBIS scheme is the content standard used by OBIS and is designed for marine biodiversity data, specifically to record the capture or observation of a

particular species at a certain location. It can also be used for documenting specimens from museum collections and literature data. The scheme lists 74 data fields, of which 7 are mandatory and an additional 15 are classified as highly recommended. All other data fields are optional. For a full overview of the OBIS scheme, one is referred to the OBIS website.

Available Webservices on EurOBIS

The EurOBIS data system provides two main sets of data services: DiGIR (Distributed Generic Information Retrieval) which is a protocol for a distributed database system. Joining EurOBIS through this distributed data system as a contributor is relatively easy, as no adaptations to the local database structure are necessary: contributors only have to establish a link between the EurOBIS data fields and their own data fields, allowing DiGIR to recognise compliant fields and information.

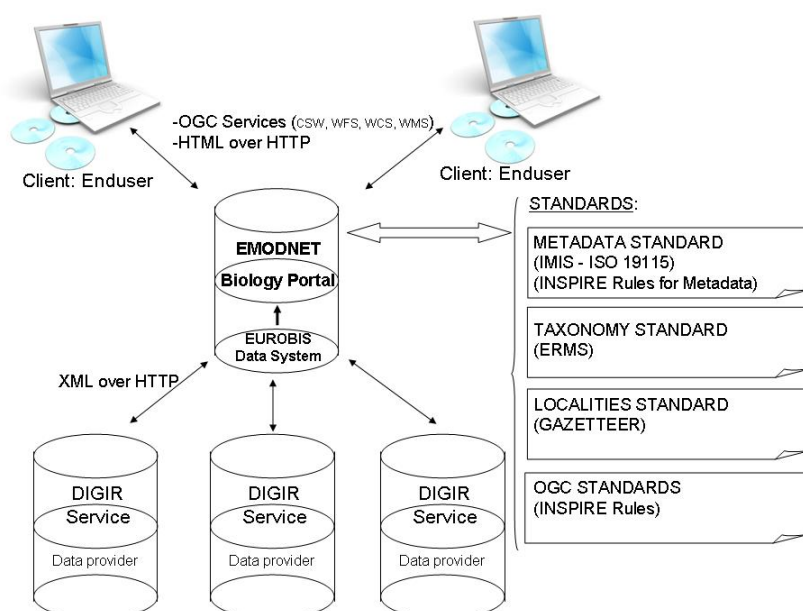


Figure: system architecture of EurOBIS

Secondly, EurOBIS serves its data and metadata also as OGC compliant webservices. The usage of these OGC Standards allows to serve geographic data and metadata as a WMS (Web Map Service) to serve geographic maps and WFS (Web Feature Service) to serve geographic vector data.

Also EMODnet, the European Marine Observation and Data Network, makes use of the OGC compliant webservices to communicate and interchange marine environmental data layers. As the Portal for Marine Ecological GenomiX (megx), a web portal for specialized georeferenced databases and tools for the analysis of marine bacterial, archaeal, and phage genomes and metagenomes also makes use of the same OGC standards, easy interoperability between EurOBIS, WoRMS, megx and other environmental data can be established. Figure 1 shows the geographic distribution of prokaryote genomes, and the following data attributes: 1) location details, 2) coordinates, 3) Date, 4) Depth, 5) habitat, 6) interpolated environmental data and 7) link to more environmental data stored at

PANGAEA, served from the megx portal and displayed on the EMODnet biology portal. All distributions of the same species *Nodularia spumigena*, accessible through EuroBIS are also visualised on the portal.

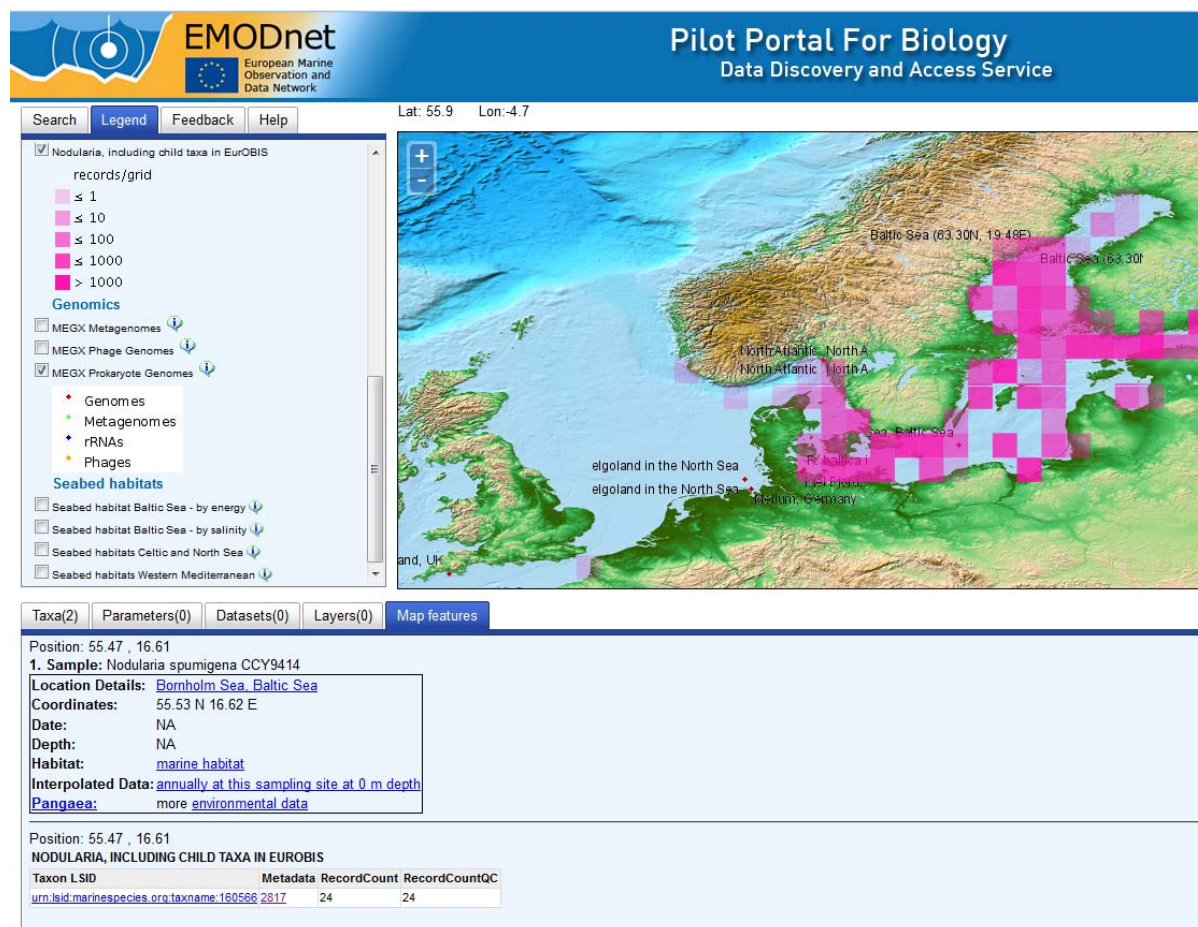


Figure: Example showing the integration of WoRMS, EuroBIS, EMODnet and megx data systems, using OGC webservices.

EuroBIS stores various millions of species observations and specimen collections. It is supported by the World Register of Marine Species (WoRMS), a standard list of > 250,000 species names and > 300,000 taxa.

EMODnet Biology portal

At this moment the EMODnet Biological data portal provides data access to 386 datasets or data collections representing 13,622,567 records of species of phytoplankton, zooplankton, angiosperms, macro-algae, benthos, birds, mammals, reptiles and fish from all European seas. The portal provides metadata on more than 600 marine biological datasets and gives access to 374 data products. Through the involvement of data providers, both managers of large thematic databases and organizations coordinating national marine biological monitoring programmes, it is planned in the next phase of EMODnet Biology to mobilize more long-term biological monitoring data. Extra thematic databases that will provide direct access to EMODnet include for example long-term near-surface phytoplankton and



zooplankton data from the North Sea and North East Atlantic collected by the Continuous Plankton Recorder (CPR) , ship-based and aerial observations of birds and mammals from the European Seabirds at Sea database, fish data from the International Bottom Trawl Survey database and several long-term national marine biological monitoring databases from different European countries. Also activities are foreseen to mobilize historical datasets into the system, relevant for creating baseline information for comparison with current and future assessments of the marine environment.

Possible storage of environmental clades and Operational Taxonomic Units (OTU) in EurOBIS

This is analyzed using the OSD sampling as a case study. VLIZ is also looking into the activities carried out during the MICRO project. The MICROBIS database (<http://www.marbef.org/data/imis.php?module=dataset&dased=1980>) serves legacy, lipidomic, and pyrosequencing data and associated contextual data collected as part of the ICoMM Census of Marine Life Ocean Realm Project. Legacy data include geospatial and environmental data collected from environmental sequencing surveys prior to the initiation of the ICoMM effort. The geospatial data from MICROBIS are stored in EurOBIS and can act as case study for storing the OSD biogeographic data.

WoRMS - The World Register of Marine Species

The World Register of Marine Species is also managed by VLIZ and it provides an open-access inventory of all marine species. It is 90% complete, and continually expands its content beyond species names and synonyms. WoRMS is an authoritative taxonomic list of species occurring worldwide in the marine environment. The taxonomic register is being used for taxonomic efficiency and quality control in marine biodiversity research and management.

As part of Micro B3 VLIZ has checked specifically the content of WoRMS in terms of lower organisms, and the fitness for purpose for microbial organisms with the following results:

A. Review of Bacteria & Archaea in WoRMS (work started in 2011)

Those groups, specially the Bacteria, include an enormous number of taxa and go through constant changes due to the on-going research and permanent innovation. Therefore, a protocol to proceed with the review had to be chosen:

1. Search the National Center for Biotechnology Information (U.S. National Library of Medicine) Taxonomy Page – GenBank (<http://www.ncbi.nlm.nih.gov/>).
2. Make a list up to Genus level with all the accepted names.
3. Comparing the list with the WoRMS data base to estimate the further work.
4. Search in OBIS, Ocean biogeographic Information System (<http://iobis.org/>) to list



the genera, and species when mentioned, with a marine distribution.

5. A revision of all the species (with a name in the traditional genus+sp) listed in GenBank was performed.
6. Consulting the papers and publication references if available. All those species with a clear marine origin were added to the Aphia database.
7. When possible, an authority, one or more publication references, distribution of the species, possible host and/or feeding type was mentioned.

This way a total of more than 1000 new entries and corrections were made to prepare WoRMS for use in Micro B3..

The main problems when trying to systematize a group such as Archae or Bacteria is the uncertainty that still exists around them: many cases of strains identified but not properly studied, published and named (only an ID code, unclassified samples, uncultured samples, environmental samples); constant changes in the taxon tree as new genetic analyses are performed. On the other hand, and concerning WoRMS, it is not always clear when to accept a species as marine. Connection between land and sea, through human activity, waste waters, etc increase the presence of usually terrestrial or animal/human parasite bacteria in the sea but that does not mean that they are necessarily marine living forms.

VLIZ matched the prokaryotes of WoRMS also with the SILVA database (A comprehensive on-line resource for quality checked and aligned ribosomal RNA sequence data, <http://www.arb-silva.de/>). Taxonomic path and taxa name matches were performed for the WoRMS data on the SILVA data. It was found that WoRMS has 1816 distinct taxa, and SILVA has 3179. When these were matched on a path-to-path basis, there are 1005 in common, 2174 only in SILVA, and 811 only in WoRMS. When a taxa name matching was performed, the situation improves and 1595 names are shared. But this also implies that, some paths are different from WoRMS, or vice versa (could be spelling, different higher taxa rank names used etc..).

B. Review of viruses in WoRMS (work started in 2012)

The names of the Orders, Families, Subfamilies and Genera of viruses, have been added in the World Register in May 2012, based on :

King, A.M.Q.; Adams, M.J.; Carstens, E.B.; Lefkowitz, E.J. (Ed.) (2012). Virus taxonomy: classification and nomenclature of viruses. Ninth report of the International Committee on Taxonomy of Viruses. Elsevier Academic Press: Amsterdam. ISBN 978-0-12-384684-6. x, 1327 pp.

Both the higher taxonomic hierarchy and the species pose some difficulties:

1. The higher taxonomy (above Order) doesn't have a formal taxonomic rank, nor a



scientific name. Orders and Families are grouped together based on genome composition and structure (DNA viruses with groups dsDNA and ssDNA viruses, RNA viruses with groups dsRNA, negative sense ssRNA and positive sense ssRNA viruses, ...). Within WoRMS, these groups have currently been added as Phyla and Subphyla, although this is not completely according to the International Nomenclature Code.

2. The name giving of the species does not follow the Code of Nomenclature. Species names are usually derived from the common name of the virus (usually in English) used to establish the species. Examples : Elephantid herpesvirus 1, Dicliptera yellow mottle virus, human parainfluenza virus 3, ... VLIZ still has to evaluate how these names can be added to the World Register of Marine Species as they do not completely coincide with the currently followed Nomenclature Code.
3. It is not always evident to separate the marine viruses from the non-marine. Some information seems to be available, and it will mainly be a manual task to feed this information into WoRMS.

C. Review of protists in WoRMS

WoRMS classification currently Catalogue of Life. VLIZ is looking at the publication: The New Higher Level Classification of Eukaryotes with Emphasis on the Taxonomy of Protists (Adl et al). The main issue is that Adl et al contains unranked classes. <http://onlinelibrary.wiley.com/doi/10.1111/j.1550-7408.2005.00053.x/abstract>.

D. Life Science Identifier (LSID)

LSIDs are persistent, location-independent, resource identifiers for uniquely naming biologically significant resources. The LSID concept introduces a straightforward approach to naming and identifying data resources stored in multiple, distributed data stores in a manner that overcomes the limitations of naming schemes in use today. WoRMS has implemented LSIDs for all its taxonomic names and they are displayed on each taxon page. VLIZ has integrated the AphiaID into the LSID, so in fact you can continue using the AphiaID. These LSID's are key identifiers in linking marine microbial and biological data systems.

E. Available Webservices on WoRMS

The Bioinformatics Data Systems can use the SOAP / WSDL webservice built on the World Register of Marine Species. WoRMS uses the platform independent SOAP/WSDL standard. SOAP (Simple Object Access Protocol) is a computer protocol that can be used to communicate between two datasytem, using the World Wide Web Hypertext Transfer Protocol (HTTP) and the Extensible Markup Language (XML), defined through the Web Service Definition Language (WSDL). Possible information that can be requested through the webservice is:

- get the AphiaID for your taxon
- check the spelling of your taxa



- get the authority for your taxa
- get the full classification for your taxa
- resolve your unaccepted names to accepted ones
- get all synonyms for a taxon
- match your species list
- resolve a common name/vernacular to a scientific name
- get the common name(s)/vernacular(s) for a taxon
- get the sources/references for a taxon
- get the WoRMS citation for a taxon
- get the direct children for a taxon

2.4 ICES

The International Council for the Exploration of the Seas (ICES), through its work to provide best available scientific knowledge and advice to the North East Atlantic and adjacent area, have a well established Data Centre. The ICES Data Centre manages a number of large dataset collections related to the marine environment covering the NE Atlantic, Baltic Sea, Greenland Sea and Norwegian Sea. The data originate from national institutes that are part of the ICES network of member countries that include all countries bordering the NE Atlantic and Baltic Seas.

The data are primarily related to national monitoring programmes and associated to a number of thematic areas including Fisheries, environmental pollution and effects, biodiversity and the physical conditions of the sea and seabed. The Data Centre holds contracts with the regional sea conventions (HELCOM and OSPAR), as well as the European Environment Agency (EEA) to manage marine datasets on their behalf. Data coming into the ICES dataset collections undergoes a number of automated checks including checks on format, required information, range checks, valid references, outliers and cross-references. In addition a number of visual checks are made by the data managers, before the data is released to the data portals. Because the data portals are specifically used for a number of regional assessments related to the Regional Sea Conventions and the Data Collection Framework, a continuous check and feedback on data are made by these users.

ICES organises these dataset collections around specific thematic data portals, as well as an overarching data warehouse. The table below is a summary of the information that can be visualised and downloaded in the data warehouse <http://ecosystemdata.ices.dk>

Dataset Collection	No of Measurements	of No of years	No of Parameters/Taxa
Oceanographic	239,470,116	123	18
Contaminants and biological effects	10,016,078	35	646



Fish trawl survey	5,296,448	48	489
Fish predation (stomach contents)	1,149,608	12	845
Biological community	690,251	33	2,200
ICES Historical Plankton	318,319	11	1,985

Data portals

The dataset collections are organised around a number of thematic data portals, which are described below.

Oceanographic data is made available through the web applications at <http://ocean.ices.dk/>. The different applications are a mixture of ‘on-the-fly’ queries to optimised web ready databases, as well as static content that is prepared in batch jobs/cached and usually runs on a nightly or once a week update.

Contaminants, biological effects and **biological community** data are made available through the web portal <http://dome.ices.dk> (Database on Oceanography and Marine Ecosystems). This portal has a map visualisation function using the Google API, as well as the ability to download individual dataset files. The regional extractions for OSPAR are made from this database and it is planned in Autumn 2012 to provide this functionality in the DOME portal directly.

Fish Trawl Survey datasets collected in connection with the Data Collection Framework (EU-DCF) are managed under the <http://datras.ices.dk> portal. This portal has an upload and screening facility, as well as downloads of a number of fisheries related data products.

Fish predation and **Historical plankton** are ‘historical’ dataset collections, where the dataset is considered complete and there are no immediate plans to update them. Each have a data portal <http://ecosystemdata.ices.dk/HistoricalPLankton/> and <http://ecosystemdata.ices.dk/stomachdata/> built on the same framework with the same functionality as can be found on the EcoSystemData portal.

Other portals and services

Web services have been developed around a number of generic queries, as well as some more specific applications, such as the platform code service for SeaDataNet, and the Species records for EMODNet Biology. A shortlist can be found on each web portal that has services, and also on the ‘Web services’ tab of this page

<http://www.ices.dk/datacentre/Submissions/index.aspx?t=1>

Controlled **Vocabularies** are integral to all of the ICES dataset collections and are managed and maintained through the <http://vocab.ices.dk> web page. In addition, specific applications have been made for more specialised cases, such as the SeaDataNet platform request and management system <http://www.ices.dk/datacentre/requests/>.

Where possible, ICES uses established vocabularies and uses their services to ensure minimum duplication of effort i.e. species names are managed through the World Register of Marine Species (WoRMS).

Spatial data: In addition to the specific data portal application, ICES has a spatial facility <http://geo.ices.dk> for both managing, distributing and viewing spatial data layers as well as cataloguing and discovery services for metadata related to ICES datasets and ICES working group and client data products. This portal is built on the open source geoserver and geonetwork architecture.

Technology platform

The main database technology for ICES databases is Microsoft SQL Server, with versions spanning 2005, 2008 and 2012. The main programming languages for the applications are Microsoft Visual C# .net and Microsoft Visual Basic .net. In addition the services and applications require that we also programme a variety of other languages/formats including Java, Ajax, PostgreSQL, XML, JavaScript etc.


ICES databases are managed in-house and the operating platforms used are a mixture of clustered physical server machines but increasingly virtual machines housed on clustered host servers.



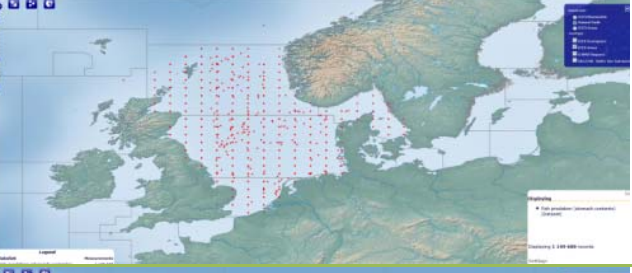


Data policy

ICES has a well established and open access minded data policy <http://www.ices.dk/datacentre/datapolicy.asp>. ICES expect that users of the data can download or gain access to the data by just accepting the policy (usually by a mouse click) and no registration or questions are levied at the user. In return, the user is obliged to duly acknowledge the dataset source and also report back to ICES should they discover potential issues with the data.

A review of the current data policy is expected to be completed by Autumn 2012, where more specific guidance will be added on metadata and redistribution of data.

Dataset Distribution and Metadata links

Dataset Collection	ISO Metadata record	Geographic Distribution of records
Oceanographic	Meta data link	

Dataset Collection	ISO Metadata record	Geographic Distribution of records
Contaminants and biological effects	Metadata link	
Fish trawl survey	Metadata link	
Fish predation (stomach contents)	Metadata link	
Biological community	Metadata link	
ICES Historical Plankton	Metadata link	

As part of the SeaDataNet II project it is planned that ICES will be connected soon to the SeaDataNet infrastructure for giving overview and access especially to its oceanographic and contaminants data sets via the SeaDataNet CDI Data Discovery and Access service. ICES maintains an international database and considerable overlap will exist with the collections of the SeaDataNet national data centres (NODCs). Therefore careful attention will be given to selecting complementary data sets that will be made accessible via the SeaDataNet portal.



ICES is also a biological data contributor to EurOBIS and the EMODNet Biology portal, which is based upon EurOBIS. This will be continued and expanded in the EMODNet Biology 2 project that has recently been proposed to EU DG MARE and which has a good chance of getting awarded.

2.5 PANGAEA

PANGAEA®; (Data Publisher for Earth & Environmental Science, www.pangaea.de) provides essential services like scientific project data management, long-term data archiving, data publication, and dissemination via visualisation and analysis software (freeware products) and via interoperability according to international protocols and standards. PANGAEA® holds mandates from ICSU (World Data Centre for Marine Environmental Sciences - WDC-MARE) and WMO (World Radiation Monitoring Center - WRMC).

PANGAEA manages and can provide environmental and biological data, including an extensive range of parameters describing the life history and vital rates of marine plankton (viruses, bacteria, autotrophic and heterotrophic protists, crustaceans and jellyfish) and microbenthos from contemporary to paleobiogeographic records. Furthermore, PANGAEA publishes and archives a large volume of data sets from scientific projects such as HERMES, EUROCEANS, HERMIONE, SESAME, and ESONET, EURO-BASIN, JGOFS.

Datasets published by PANGAEA® are extremely diverse and include for example water column profiles, sediment core profiles, biogeographic distributions, meteorological and oceanographic time series, vital rate measurements on individual organisms or communities, sea bed photographs, audio and video records, and ROV surveys. PANGAEA® integrates biological data on plankton, fish, corals, micro-, meio- and macrobenthos, marine mammals and birds. In addition to contemporary observations it offers an extensive paleobiogeographic record from marine sediment cores. Biological data include key biodiversity parameters such as taxon-specific abundance and biomass, and also an extensive range of parameters describing the life history and vital rates of marine life. At present PANGAEA® holds metadata and data for more than 600.000 data sets comprising more than 5 billion data items for more than 80.000 different measurement types (parameters), more than 17.000 principal investigators (authors), and around 11.000 references to journal articles. Some ongoing projects are under moratorium so that access to data may be restricted, but metadata is always freely accessible.

PANGAEA® - technical architecture

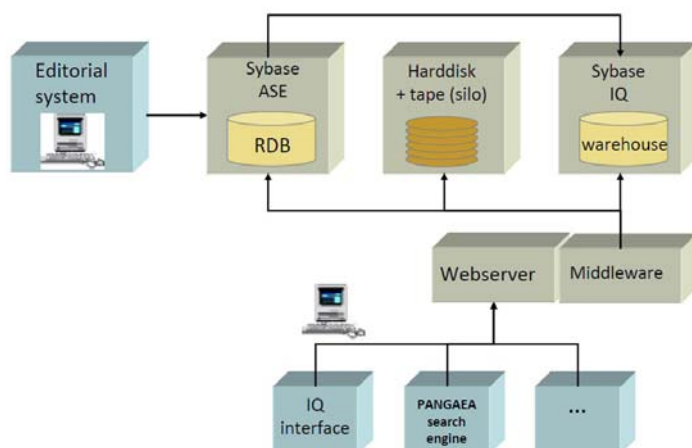


Figure: technical architecture of PANGAEA system

PANGAEA® has a number of standard interfaces for supporting and exchanging metadata as indicated in the following figure.

PANGAEA – standard interfaces for metadata

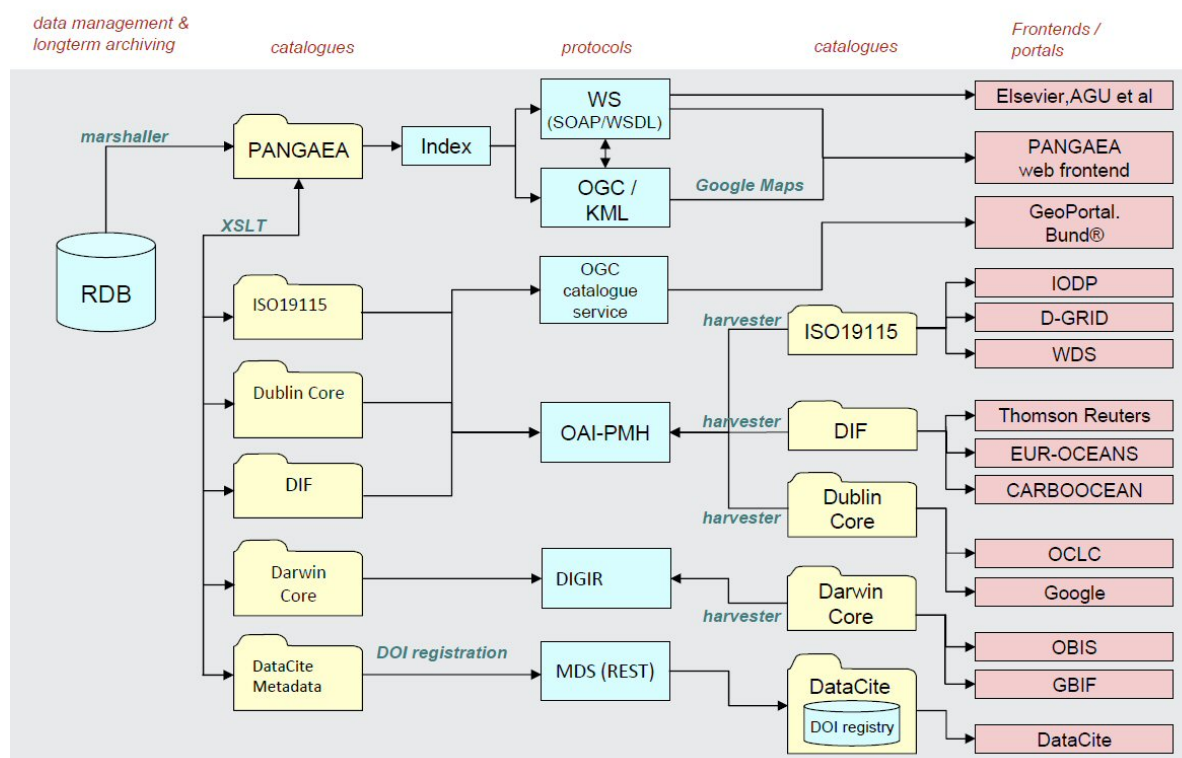


Figure: PANGAEA standard interfaces for metadata

As part of Micro B3 UniHB will lead Task 3.3 to support data management for the **Tara Ocean** expedition cruise with support of VLIZ. They will ensure that the Tara Oceans cruise data become part of the oceanographic data infrastructure for Micro B3 use and wider use.



The latter will be implemented by storing Tara Oceans metadata and data in the PANGAEA system and by arranging that the genomic samples will be forwarded to EMBL-EBI for sequencing and storing of sequences.

Furthermore as part of the SeaDataNet II project it is planned that PANGAEA will be connected soon to the SeaDataNet infrastructure for giving overview and access to its data sets via the SeaDataNet CDI Data Discovery and Access service. This CDI metadata population will be done in a gradual way and with a focus on oceanographic environmental data sets such as CTD profiles, water quality samples, geological cores etc. Thereby extra attention will be given to possible duplicates because PANGAEA is an international database and other copies of data sets might already been stored in NODCs and that way included in the SeaDataNet index.

PANGAEA is also a biological data contributor to EurOBIS and the EMODNet Biology portal, which is based upon EurOBIS. This will be continued and expanded in the EMODNet Biology 2 project that has recently been proposed to EU DG MARE and which has a good chance of getting awarded.