Marine

Microbial Biodiversity,

Bioinformatics & Biotechnology

Grant agreement n°287589

Acronym: Micro B3

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

Deliverable 3.6

Delivering dedicated data sets for Micro B3 through the oceanographic services

Version: 2.0

Circulated to: WP3 partners

Approved by: Prof. Frank Oliver Glöckner, 09/10/2015

Expected Submission Date: 30/06/2014

Actual Submission Date: 09/10/2015

Lead Party for Deliverable: IFREMER

Mail: catherine.borremans@ifremer.fr   **Tel.: +33 (0)2 98 22 41 90**

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

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. This deliverable D3.6 describes more particularly the harvesting process for recently reviewed selected oceanographic sites (see D3.4 update) in order to provide access to a first set of oceanographic data through SeaDataNet services to Micro B3 Information System. This is possible through extension of the SeaDataNet infrastructure with modules for configuring and managing dynamic and internal central buffers of data sets at SeaDataNet fitting the Micro B3 purpose (see D3.5).

After selection on D3.4 sites extra data sets were identified for their relevance in the Micro B3 context but missing in the SeaDataNet CDI service. Therefore arrangement with data providers to populate SeaDataNet was made, especially for the GOSUD project data sets. The data provision from SeaDataNet to Micro B3 concerns unrestricted and under SeaDataNet license data sets, which have been and are automatically harvested. Therefore related data centres were contacted to make them aware of the Micro B3 interest to make use of their data sets for filling the Micro B3 buffer.

Based on the final list of sites of interest a Micro B3 filter was then set up on the SeaDataNet CDI database thanks to the robot harvesting system developed by MARIS. The buffer of harvested data is accessable for the Micro B3 Information System via a controlled API, but the results of the Micro B3 filter can also be seen by a dedicated version of the public CDI Data Discovery and Access service interface: http://mb3.maris2.nl/v_cdi_v3/search.asp. At the time of writing this deliverable, this represents almost 47000 records divided between 134 originators from 21 different countries and covering 34 parameter groups.

This compilation and delivery of oceanographic data sets will allow to integrate biological information on the diversity and function of marine microorganisms with the environmental conditions surrounding them. This will in turn offer new perspectives for the modelling and exploration of marine microbial communities.

Table of Content

Introduction

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. Then, the overarching objective of Micro B3 is to provide access and integrate genomic, oceanographic and earth observation databases into one Micro B3 Information System (Micro B3-IS), based on global standards for sampling and data processing.

Work and deliverables achieved through WP3 (in close collaboration with WPs 2, 4, 5 and 6) have led to the present Deliverable 3.6: the actual gathering and delivery of dedicated data sets for Micro B3 through the oceanographic services.

Indeed, a model for organising data flows was initially agreed between oceanographic environmental (SeaDataNet), marine biodiversity (EurOBIS) and genomic (ENA) data infrastructures in order to ensure that the data needs for Micro B3 will be served, at least for data originating from European in-situ sources.

Additional oceanographic data sets collected in the framework of Micro B3 via the Ocean Sampling Day (OSD) (WP2) and derived from the Tara Oceans expedition are managed by PANGAEA through which SeaDataNet is further populated with these data. Genomic samples are transferred to EMBL-EBI for processing and inclusion in its European Nucleotide Archive (ENA). ICES and Malaspina expedition data are integrated in the CDI database as well.

Horizontal interoperability between SeaDataNet, EurOBIS and ENA was planned for all marine data on a global scale while vertical interoperability of these European systems towards Micro B3-IS was arranged for a specific subset of the data related to geographical pilot areas defined in D3.4. This architecture will result in an improved and complete data provision to internal and external users as they will have direct overview of data sets as managed and provided by each of these infrastructures.

Deliverable D3.5 ("Portal, supporting OGC and ISO standards, bundling the selected oceanographic services for Micro B3" publicly available at the link https://www.microb3.eu/sites/default/files/deliverables/MB3_D3_5_PU.pdf) describes the further detailing, development and actual implementation of the selected interoperability solutions for sharing metadata and data from SeaDataNet towards MB3-IS, EurOBIS and ENA. The implemented interoperability solutions concern machine-to-machine services for sharing metadata and data from SeaDataNet towards MB3-IS, and OGC WMS-WFS services as well as OpenSearch services for sharing metadata from SeaDataNet towards EurOBIS and ENA.

This deliverable D3.6 describes more particularly the harvesting process for the selected oceanographic sites in order to provide access to a first set of oceanographic data through SeaDataNet services to Micro B3-IS. EurOBIS and ENA are central databases and provide direct access facilities for Micro B3-IS through their interfaces. SeaDataNet is a distributed

infrastructure whereby more than 100 data centres are connected. They share a common discovery service and shopping mechanism to find and retrieve data sets from their distributed databases. Therefore the SeaDataNet harvesting process was developed to compile dynamically a central buffer of oceanographic data sets, that can be accessed by the Micro B3-IS via a dedicated API (as described in D3.5).

Preparing actual harvesting

The objective of D3.4 was to propose a selection of pilot areas to test portal and services for gathering and making available data sets and products from the environmental data infrastructures, information that will be available for the Micro B3-IS and will allow to answer the Micro B3 Use Cases (defined by WP4, see D4.1). A first version of D3.4 based on the OSD sampling sites registry (at that time), the long term sampling sites and the Tara Oceans and Malaspina expeditions stations was submitted in February 2013.

In October 2014 it was proposed and agreed to review D3.4 in order to consider all the unique stations sampled during the first OSD event (on June 21st, 2014) because of their relevance for the Micro B3 Use Cases. In May 2015 an updated version of the Deliverable D3.4 with selected oceanographic sites was consequently prepared and submitted.

Further population of the SeaDataNet CDI database

After selection on D3.4 sites extra data sets were identified for their relevance in the Micro B3 context but missing in the SeaDataNet CDI service. Therefore arrangement with data providers to populate SeaDataNet was made, especially for the GOSUD project data sets: 132 CDI entries were generated for data from French and foreign vessels.

Further creation of SeaDataNet CDIs for significant data sets will be encouraged and supported until the end of the project and beyond.

Informing related data centres

The developments for the machine-to-machine interoperability have focused on extending the SeaDataNet infrastructure with modules for configuring and managing dynamic and internal central buffers of data sets at SeaDataNet such as fitting the Micro B3 purpose, and setting up central SeaDataNet services for giving access to the metadata and data sets in the buffers for providing MB3-IS selected data sets (see D3.5). New and updated CDI submissions will be dynamically followed by the system so that, if data are modified, the buffer for Micro B3-IS will be updated as well.

Testing has taken place to assure that the robot harvesting and dynamic maintenance of the central buffers are performing in a precise way and respecting possible data access restrictions. Not all data sets in SeaDataNet are unrestricted or accessable under the SeaDataNet license, but might require negotiations. For Micro B3 it was decided to focus

only on unrestricted and accessible under the SeaDataNet license data sets, because that will facilitate further public use in the MB3-IS.

Unrestricted or under SeaDataNet license data sets can be automatically harvested by the robot service and require no further action from their data centres. However to keep goodwill and to share interest, related data centres were contacted to make them aware of the Micro B3 activity for building the Micro B3 buffer and using these data in the MB3-IS. 36 data providers were contacted (see table below), 15 of them acknowledged receipt and 21 did not answer. Most importantly, no objection was expressed by any of the data providers.

CDI Partner	Country	Feedback
Aarhus University, Department of Bioscience, Marine Ecology Roskilde	Denmark	No answer
All-Russia Research Institute of Hydrometeorological Information (RIHMI-WDC)	Russian Federation	No answer
British Geological Survey, Edinburgh	United Kingdom	No answer
British Oceanographic Data Centre	United Kingdom	OK
Bulgarian National Oceanographic Data Centre(BGODC), Institute of Oceanology	Bulgaria	No answer
ENEA Centro Ricerche Ambiente Marino - La Spezia	Italy	No answer
Finnish Meteorological Institute	Finland	OK
Flanders Marine Institute	Belgium	OK
Geological Survey of Denmark and Greenland (GEUS)	Denmark	No answer
German Oceanographic Datacentre (NODC)	Germany	No answer
Hellenic Centre for Marine Research, Hellenic National Oceanographic Data Centre (HCMR/HNODC)	Greece	OK
IEO/Spanish Oceanographic Institute	Spain	OK
IFREMER / IDM / SISMER - Scientific Information Systems for the SEA	France	OK
IHPT, Hydrographic Institute	Portugal	OK
Institute of Biology of the Southern Seas, NAS of Ukraine	Ukraine	No answer
Institute of Fishery Resources (IFR)	Bulgaria	No answer
Institute of Marine Research - Norwegian Marine Data Centre (NMD)	Norway	No answer
Institute of Oceanography and Fisheries	Croatia	No answer
International Council for the Exploration of the Sea (ICES)	Denmark	OK
ISPRA-Institute for Environmental Protection and Research	Italy	No answer
Israel Oceanographic and Limnological Research (IOLR)	Israel	No answer
Management Unit of North Sea and Scheldt Estuary Mathematical Models, Belgian Marine Data Centre	Belgium	OK
Marine branch of Ukrainian Hydrometeorological Institute	Ukraine	No answer
Marine Hydrophysical Institute	Ukraine	No answer

CDI Partner	Country	Feedback
Marine Institute	Ireland	OK
Marine Research Institute	Iceland	No answer
National Institute of Fisheries Research (INRH)	Morocco	No answer
National Institute of Meteorology and Hydrology, Bulgarian Academy of Sciences	Bulgaria	No answer
NIOZ Royal Netherlands Institute for Sea Research	Netherlands	OK
OGS (Istituto Nazionale di Oceanografia e di Geofisica Sperimentale), Division of Oceanography	Italy	OK
PANGAEA - Data Publisher for Earth & Environmental Science	Germany	OK
Portuguese Institute of Ocean and Atmosphere	Portugal	OK
SHOM (SERVICE HYDROGRAPHIQUE ET OCEANOGRAPHIQUE DE LA MARINE)	France	No answer
Swedish Meteorological and Hydrological Institute	Sweden	OK
TNO Geological Survey of the Netherlands	Netherlands	No answer
Ukrainian scientific center of Ecology of Sea (UkrSCES)	Ukraine	No answer

Delivering dedicated data sets

Interfacing from SeaDataNet towards Micro B3-IS

As briefly introduced above, the SeaDataNet CDI system is used for giving overview and access by means of a shopping mechanism to the distributed data sets that are managed by connected data centres. To deal with the requirements for Micro-B3 as well as for a number of other user communities a robot harvesting system has been developed by MARIS for automatic and more efficient discovery and harvesting of metadata and data sets that might be transferred or made accessible for applications of specific user communities. This robot buffer system is used in the Micro B3 project context in order to perform the dynamic production harvesting of SeaDataNet data for the Micro B3-IS selected geographic sites.

An Application Programming Interface (API) was provided to Micro B3-IS to create a client service to submit and retrieve selected data sets from the SeaDataNet – Micro B3 buffer. This API will work next to the Central Buffer User Interface and will facilitate remote and authorised machines such as Micro B3-IS to logon, using the SeaDataNet AAA service, and then to perform 3 functions: Create_order, Get_order_info, Download_order.

These systems were largely described in Deliverable D3.5.

Actual harvesting

Based on the final list of sites of interest a Micro B3 filter was then set up on the SeaDataNet CDI database. The buffer of harvested data is accessible for the Micro B3-IS via a controlled API, but the results of the Micro B3 filter can also be seen by a dedicated version of the public CDI Data Discovery and Access service interface:
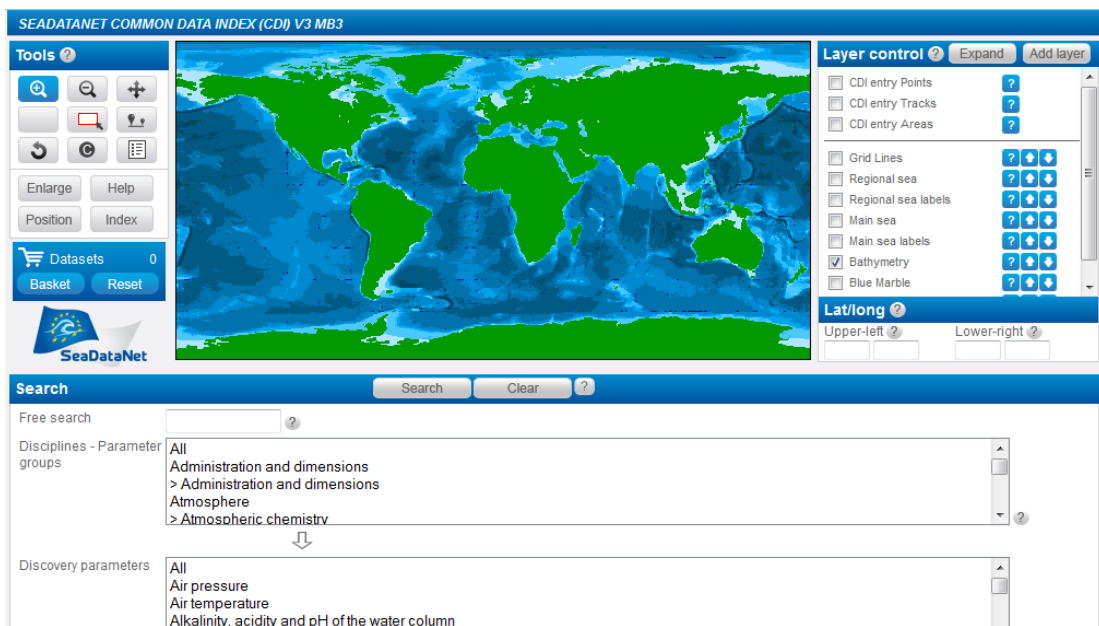
http://mb3.maris2.nl/v_cdi_v3/search.asp.

Figure 1: SeaDataNet CDI interface for the Micro B3 project purpose

When pressing the "Search" button, all data sets (unrestricted/SDN license) available for use in Micro B3-IS are displayed. At the time of writing of this deliverable, this represents almost 47000 records.
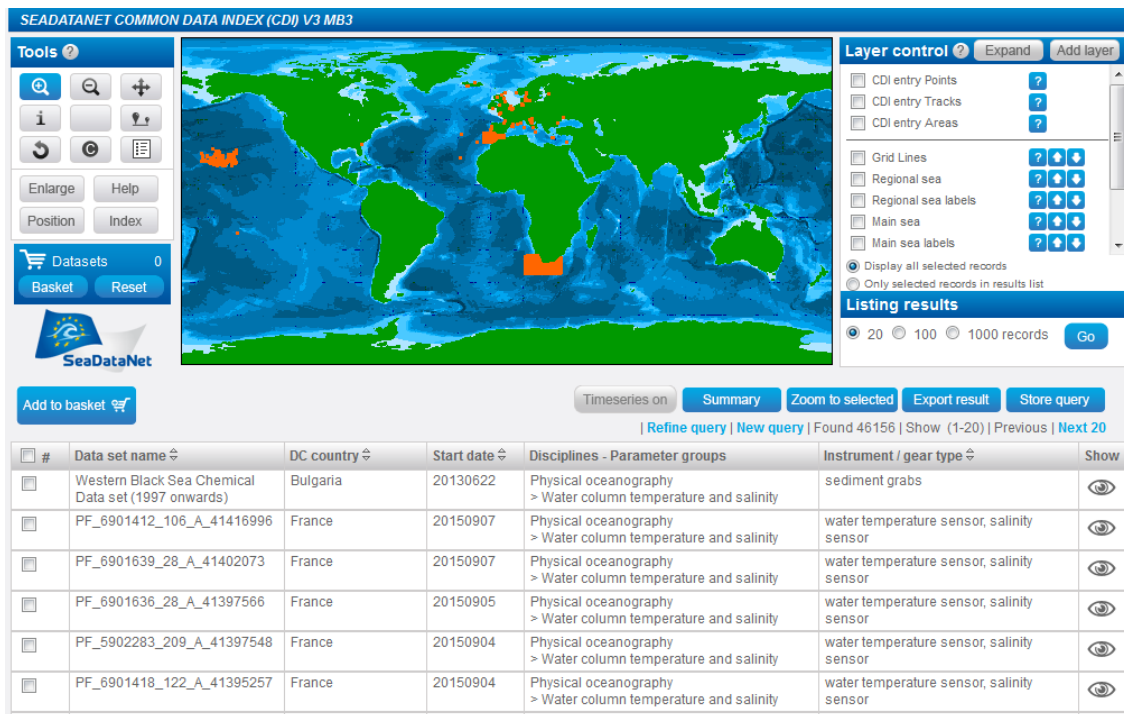

Figure 2: SeaDataNet CDI interface filtered for the Micro B3 project purpose

A radius of 0.05 degrees around point locations (mainly OSD sites) was used in order to perform the spatial query on the CDI system and search automatically which data sets are in these boxes as area of interest.

Pressing the "Summary" button gives the overview of the data centers that are involved and clicking the data center shows which data sets from each is being harvested for use in Micro B3. "Summary" button gives also an overview of results by Originator, Parameter Group, Discovery Parameter, Instrument/Type, Year and Data Access Restriction.

Figure 3: Results summary by CDI Partner, Originator, Parameter Group, Discovery Parameter, Instrument/Type, Year and Data Access Restriction
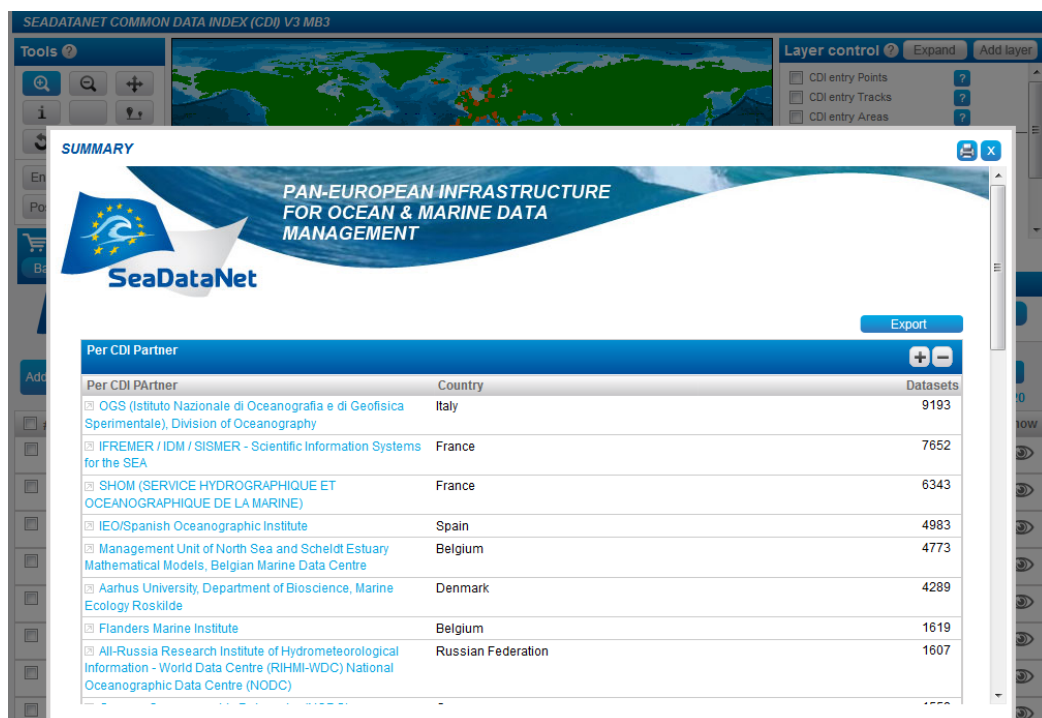

Figure 4: List of data centers related to the Micro B3 filter on SeaDataNet data sets

9

For instance, the nearly 47000 entries are divided between 134 originators from 21 different countries.

The Micro B3 SeadataNet filter includes data related to the 34 following parameter groups:

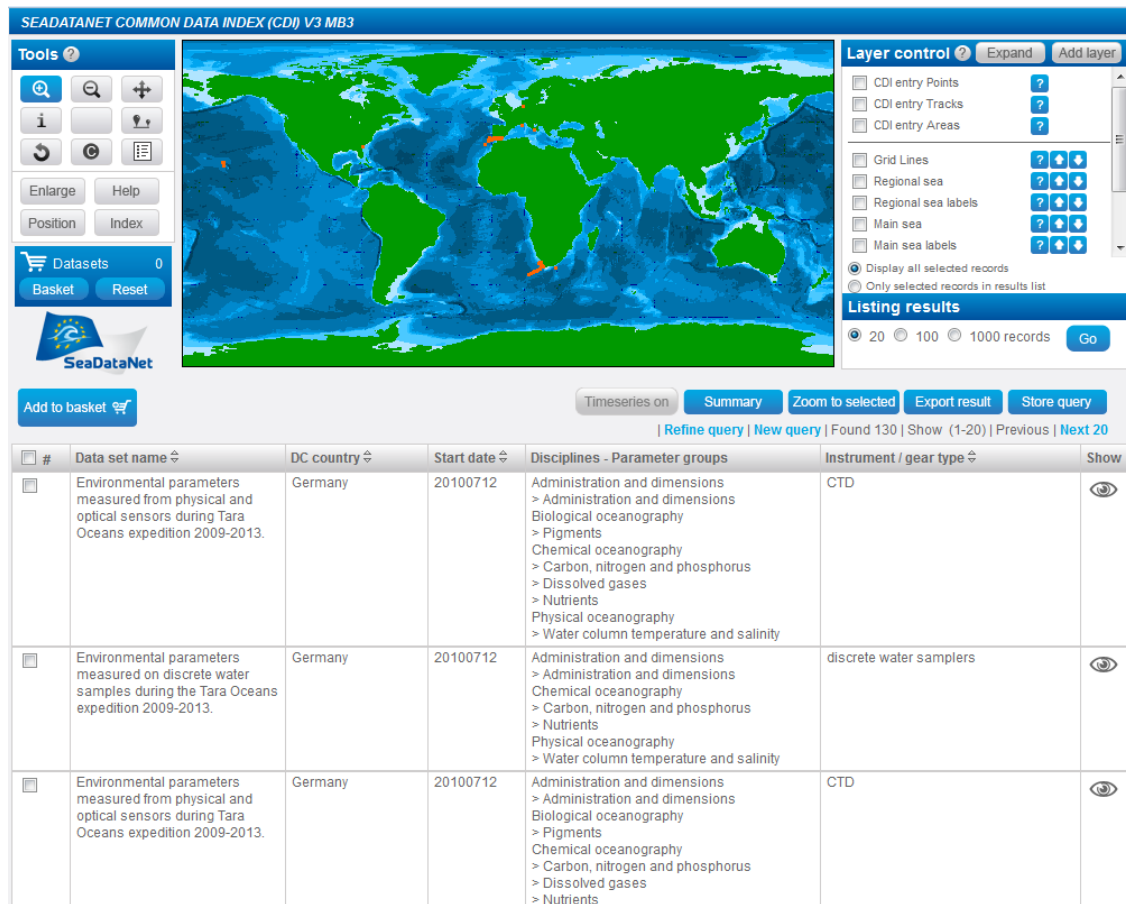| |
|---|
| Water column temperature and salinity |
| Administration and dimensions |
| Dissolved gases |
| Carbonate system |
| Carbon, nitrogen and phosphorus |
| Nutrients |
| Pigments |
| Suspended particulate material |
| Other physical oceanographic measurements |
| Optical properties |
| Metal and metalloid concentrations |
| Rock and sediment lithology and mineralogy |
| Sea level |
| Rock and sediment biota |
| Anthropogenic contamination |
| PCBs and organic micropollutants |
| Other inorganic chemical measurements |
| Hydrocarbons |
| Rock and sediment chemistry |
| Currents |
| Rock and sediment physical properties |
| Biota composition |
| Other organic chemical measurements |
| Halocarbons (including freons) |
| Other biological measurements |
| Sediment pore water chemistry |
| Meteorology |
| Fish |
| Isotopes |
| Rate measurements (including production, excretion and grazing) |
| Acoustics |
| Atmospheric chemistry |
| Fluxes |
| Fatty acids |

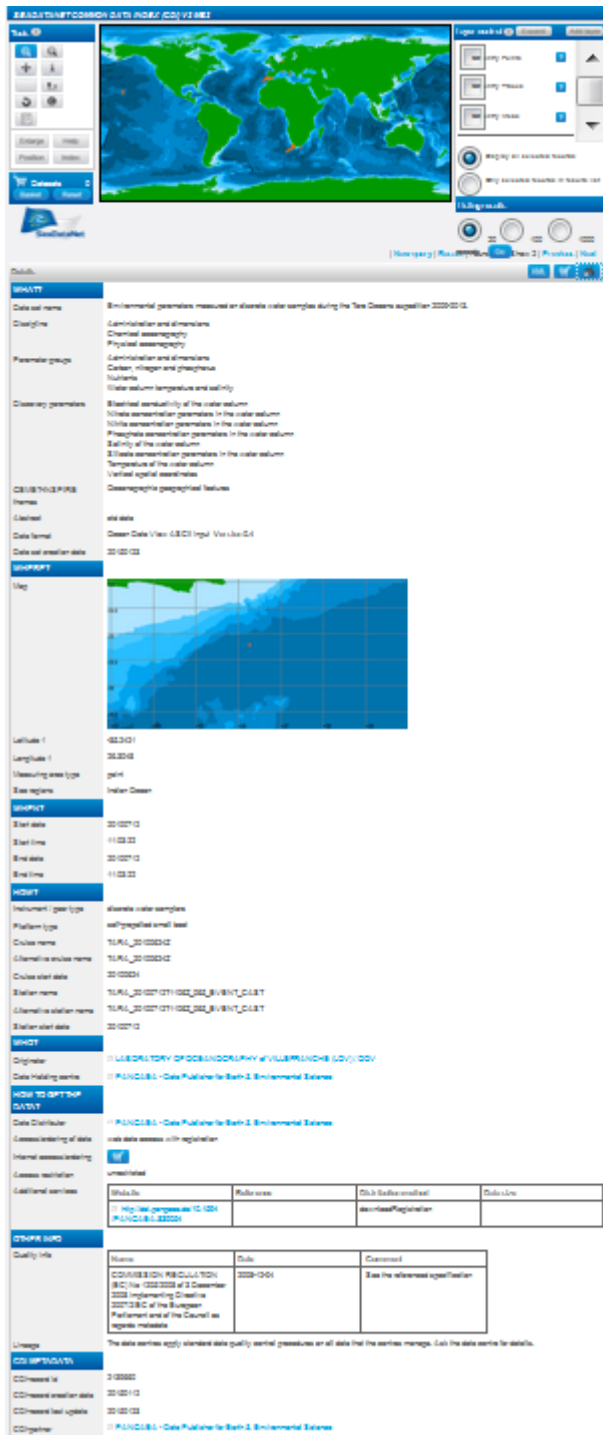**Figure 5**: List of data sets harvested for use in Micro B3 for one data center

Figure 6: CDI metadata for one data set

Conclusion

This compilation and delivery of oceanographic data sets will provide information concerning physics, chemistry, biology, geology and geophysics for the Micro B3 Use Cases. This is a prerequisite for being able to integrate biological information on the diversity and function of marine microorganisms with the environmental conditions surrounding them. These integrated datasets will lead to a better understanding of the marine ecosystem functioning and help in nailing down targets for biotechnological applications. The buffer system is dynamic and this implicates that the pool of oceanographic data sets as provided through SeaDataNet will increase and be updated over time automatically, depending on the further population of the SeaDataNet infrastructure by its connected data centres.

Annex A: Terminology

Term	Definition
AAA	Single-sign-on authentication and authorisation services to various components of the SeaDataNet infrastructure.
API	Application Programming Interface
CDI	Common Data Index metadata schema as developed and operated by SeaDataNet to describe observation data sets
ENA	European Nucleotide Archive (ENA) operated by EMBL-EBI-EMBRC for molecular sequence data
EurOBIS	European Node of the international Ocean Biogeographic Information System (OBIS) as operated by VLIZ
GOSUD	Global Ocean Surface Underway Data Project
MB3-IS	Micro B3 Information System
OGC	Open Geospatial Consortium: an international industry consortium to develop community adopted standards to "geo-enable" the Web
OpenSearch	A collection of technologies that allow publishing of search results in a format suitable for syndication and aggregation
OSD	Ocean Sampling Day
RSM	Request Status Manager, module in CDI service for administering shopping requests
SDN	SeaDataNet: EU-funded pan-European e-infrastructure for the management and delivery of marine and oceanographic data
WFS	Web Feature Service – OGC standard
WMS	Web Map Service – OGC standard