Marine Microbial Biodiversity, Bioinformatics & Biotechnology

Grant agreement n°287589

Acronym : Micro B3
Start date of project: 01/01/2012, funded for 48 month

Deliverable 3.3

Analysis and definition of interoperability of oceanographic services with Micro-B3

Version: 1.0
Circulated to: Extended Executive Board (February 2013)
Approved by: Frank Oliver Glöckner

Expected Submission Date: 31.12.2012
Actual submission Date: 15.02.2013

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 the major existing overarching oceanographic data management infrastructures, SeaDataNet and EurOBIS, that also are both driving the data management component of the EMODNet development. These oversee and give access to extensive volumes and types of data sets from existing ocean and marine data collection activities from multiple sources. Their contribution to Micro B3 will be arranged 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 gives data management support. These campaigns not only provide environmental data but also genomic samples. WP4 is drafting an **Ocean Sampling Handbook** combining input from both the genomic community and the oceanographic community. 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.

Deliverable D3.1 describes in detail the set up, scope and services of SeaDataNet, including its network of connected data centres such as NODCs, ICES, and PANGAEA, as well as of EurOBIS, and their roles in the EMODNet development. Deliverable D5.5 describes the set-up, scope and services of the EMBL-EBI's European Nucleotide Archive (ENA) and the way it will manage and deliver genomic data services to the Micro B3 Information System (MB3-IS).

This Deliverable D3.3 gives a functional analysis of the way that the flow of data from the field via the data management infrastructures to Micro B3, MB3-IS and users might be structured and organised. This has been tuned with related WPs in Micro B3: WP2, WP4 and WP5 (see also Deliverable D5.6). Furthermore this Deliverable gives a technical analysis and specifications of the interoperability options, both for delivering metadata and data to MB3-IS and for the mutual exchanges between SeaDataNet (marine environmental data), EurOBIS (marine biodiversity data) and the European Nucleotide Archive (ENA) (molecular sequence data).

The further detailing and actual implementation of the interoperability solutions, followed by the operational provision of oceanographic data to Micro B3 will take place in the second and third year of the Micro B3 project.

Table of Contents

1.0 Context

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.

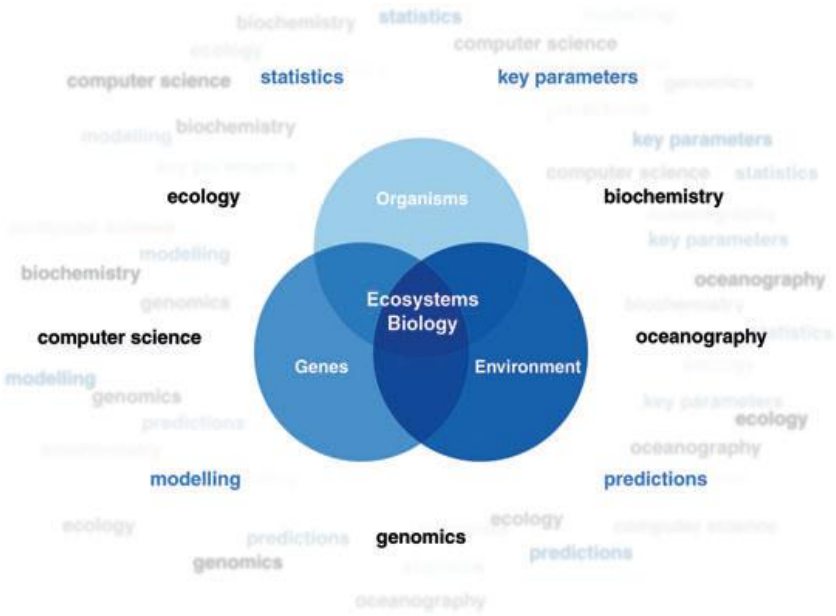

Figure: Integrating the information on the diversity of the organisms with their potential functions, reflected in their genes, and the environmental conditions surrounding them.

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.

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. 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 through WP3 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 is described in Deliverable D3.2.

One of the major challenges in Micro B3 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 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.

Deliverable D3.1 describes in detail the set up, scope and services of the leading European oceanographic data management infrastructure SeaDataNet, including its network of connected data centres such as NODCs, ICES, and PANGAEA, as well as of EurOBIS as leading European infradtrcture for marine biodiversity, and their roles in the EMODNet development. Deliverable D5.5 describes the set-up, scope and services of the EMBL-EBI's European Nucleotide Archive (ENA) and the way it will manage and deliver genomic data services to the Micro B3 Information System (MB3-IS).

WP3 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 transfering 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.

This Deliverable D3.3 has been drafted as part of Task 3-1: Establishing interoperability between Micro B3 and the Oceanographic Environmental data management systems. It gives a functional analysis of the way that the flow of data from the field via the data management infrastructures to Micro B3, MB3-IS and users might be structured and organised. This has been tuned with related WPs in Micro B3: WP2, WP4 and WP5 (see also

Deliverables D5.5 and D5.6). Furthermore this Deliverable gives a technical analysis and specifications of the interoperability solutions, both for delivering data to MB3-IS and for the mutual exchanges between SeaDataNet (marine environmental data), EurOBIS (marine biodiversity data) and the European Nucleotide Archive (ENA) (molecular sequence data).

2.0 Approach

The Task 3.1 activities have been undertaken by the following steps:

- analysis of the possible sources and the European oceanographic data management infrastructure for providing oceanographic environmental data to Micro B3

- analysis of a model for organising the data flow:
  - from European oceanographic data management infrastructure to Micro B3
  - between the EMBL-EBI's ENA genomics infrastructure and the European oceanographic data management infrastructure
  - from data acquisition programmes, such as Tara Oceans, to the EMBL-EBI's ENA genomics infrastructure and the European oceanographic data management infrastructure.

- analysis of possible interfacing between the different system components.

This has been undertaken by the WP3 partners, MARIS, VLIZ, UniHB, ICES and IFREMER by studying documentation, discussions during plenary and bilateral meetings with representatives of other WPs (WP2, WP4 and WP5), attending Workshops and following own existing expertise and experience.

The following meetings and workshops were attended to get a good insight in the types of data required by Micro B3 and indications how the provision might be arranged:

- plenary Micro B3 kick-off meeting in January 2012 in Bremen - Germany, attended by the WP3 representatives Dick Schaap (MARIS), Stephane Pesant (UniHB), Simon Claus (VLIZ) and Neil Holdsworth (ICES) to get a better overall understanding of Micro B3 and WPs related to WP3
- bilateral meeting between Stephane Pesant (UniHB) (WP3) and Renzo Kottmann (MPI) (WP5 - MB3-IS developer) in February 2012 in Bremen - Germany discussing MB3-IS data needs and possible technical interfacing of MB3-IS with European oceanographic data management infrastructure and possible international data sources (e.g. World Ocean Data atlas (WOD))
- WP4 meeting at EBI in Hinxton - United Kingdom in April 2012 that was attended by Stephane Pesant (UniHB) (WP3) discussing three case studies where environmental data are needed to complement genomic data
- Tara-Oceans workshop in Villefranche - France in May 2012 that was attended by Stephane Pesant (UniHB) (WP3) to determine what type of observational and

derived data will be useful to characterise the environmental/ecological context of biological/genomic samples

- MicroB3 sampling groups workshop at EBI in Hinxton - United Kingdom in July 2012 that was attended by WP3 representatives Dick Schaap (MARIS), Stephane Pesant (UniNB), and Simon Claus and Klaas Deneudt (VLIZ) and which gave a very good discussion and analysis on the expected data flow from the field observation towards the EMBL - EBI genomics data infrastructure respectively the European oceanographic data infrastructure
- bilateral meeting between Peter Thijsse (MARIS) (WP3) and Renzo Kottmann (MPI) (WP5 - MB3-IS developer) in July 2012 in Bremen - Germany discussing MB3-IS data needs and possible technical interfacing of MB3-IS with SeaDataNet and EMODNet data management infrastructures
- bilateral meeting between Dick Schaap and Peter Thijsse (MARIS) (WP3) with Michael Diepenbroek, Uwe Schindler, Andree Behnken, Stephane Pesant and Andrea Schaefer (all UniHB) (WP3) in July 2012 in Bremen - Germany discussing the steps and planning for interconnecting of the PANGAEA data centre to the European SeaDataNet infrastructure of data centres, the use of CDI (Common Data Index = SeaDataNet metadata standard) and DOI (Digital Object Identifier = bibliographic citation as used by PANGAEA to refer to data sets) and the planned data management of Tara Oceans data by PANGAEA.
- bilateral meeting between Dick Schaap (MARIS) (WP3) with Guy Cochrane and Petra ten Hoopen (EMBL-ENI) (WP4 and WP5) in November 2012 in Hinxton - United Kingdom discussing the interrelationships between the European Nucleotide Archive (ENA) infrastructure and MB3-IS respectively SeaDataNet and EurOBIS.

3.0 Summary of existing European oceanographic data management infrastructure

The overview in the previous chapter indicates that there is a need in Micro B3 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, Tara Ocean and Malaspina expedition sampling locations. These have to be retrieved from existing oceanographic monitoring and scientific observation activities. In Deliverable D3.4 the potential offer of these sources for Micro B3 has been analysed and a selection has been made from a geographical perspective which and what data sets might be relevant for the Micro B3 analyses.

The existing European oceanographic data management infrastructure is described in detail in Deliverable D3.1:

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

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. The SeaDataNet data centres work together on refining their standards and expanding their infrastructure and associated services, at present supported by the EU FP7 SeaDataNet II project.

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. This is facilitated by the **Common Data Index (CDI)** Data Discovery and Access service. 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 80 connected data centres, covering physical, geological, chemical, biological and geophysical data, and acquired in European waters and global oceans.

ICES and **PANGAEA** are both data centres, managing a large volume of international marine & oceanographic data sets of potential relevance for Micro B3. As part of Micro B3 UniHB leads Task 3.3 to support data management for the **Tara Ocean** expedition cruise. It 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.

ICES and PANGAEA are also partners in SeaDataNet II and underway with connecting to the SeaDataNet infrastructure for giving overview and access to their 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. For PANGAEA the 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. Like with ICES 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.

The **European Ocean Biogeographic Information System—EurOBIS**— aims at centralising the largely scattered biogeographical data on marine species collected by European institutions and to make these quality-controlled data freely available and easily accessible. 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.

ICES and PANGAEA are also biological data contributors 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.

Implementation of the **Marine Strategy Framework Directive (MSFD)** will be aided by an overarching **European Marine Observation and Data Network (EMODNet)**. This will consist of 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.

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 pilot EMODNet thematic portals. 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. The **EurOBIS** infrastructure provides the core for the EMODnet Biology portal.

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 the next chapter it will be analysed how these infrastructures can and will work together to provide a large range of data sets that will be fit for serving the needs of Micro B3.

4.0 Model for organising data flow

Considering all the different infrastructures for data management, ongoing and near-future developments and the different types of data, a model has been drafted for the way how the data flow can be organised and structured, both for Micro B3 needs and overall for the wider community of potential users.

This model has the following features:

- a distinction is made in 3 types of data:
  - oceanographic environmental data
  - marine biodiversity data
  - genomic data
- **SeaDataNet** will index and give central and unified access to the collections of oceanographic environmental data and data products via its infrastructure of distributed data centres (at present >80 in 30 countries); ICES and PANGAEA are underway in connecting to SeaDataNet and will populate also their oceanographic environmental data sets thereby giving full attention to possible duplicates
- **EurOBIS** will index and give central access to the marine biodiversity data sets and data products via the central European biodiversity database, that VLIZ manages and that is collated by contributions from many institutes, and also via a network of a number of distributed biological data centres, including SeaDataNet NODCs. The latter capacity is to be further developed as part of the EMODNet Biology 2 project. ICES and PANGAEA already have provided input to the central EurOBIS database and will expand their contributions as distributed data centres in EMODNet Biology 2
- **EMBL-EBI** will index and give central and unified access to the collections of genomic data and data products via its **European Nucleotide Archive (ENA).**
- SeaDataNet, EurOBIS and EMBL-EBI's ENA will thus provide the 3 basic data infrastructures for Micro B3 and in particular for MB3-IS for which specific interfaces will have to be specified and later configured.
- These 3 infrastructures will ensure that the data needs of Micro B3 will be served, at least for data originating from European in-situ sources. In addition MB3-IS will make use of selected international databases, such as the World Ocean Atlas - World Ocean Database of NOAA, and remote sensing data products. Moreover it can be considered to make use of MyOcean products.
- New data to be collected as part of the Tara Oceans expedition will be curated by UniHB and stored in PANGAEA for as far as it concerns oceanographic environmental and biological data; genomic samples will be transfered to EMBL-EBI for processing and inclusion in ENA; a comparable set-up should be adopted for the data sets originating from the Ocean Sampling Day (OSD). This should be included in the

Ocean Sampling Handbook so that OSD and future other data collectors will be instructed and guided

This leads to the following model for the Micro B3 data flow

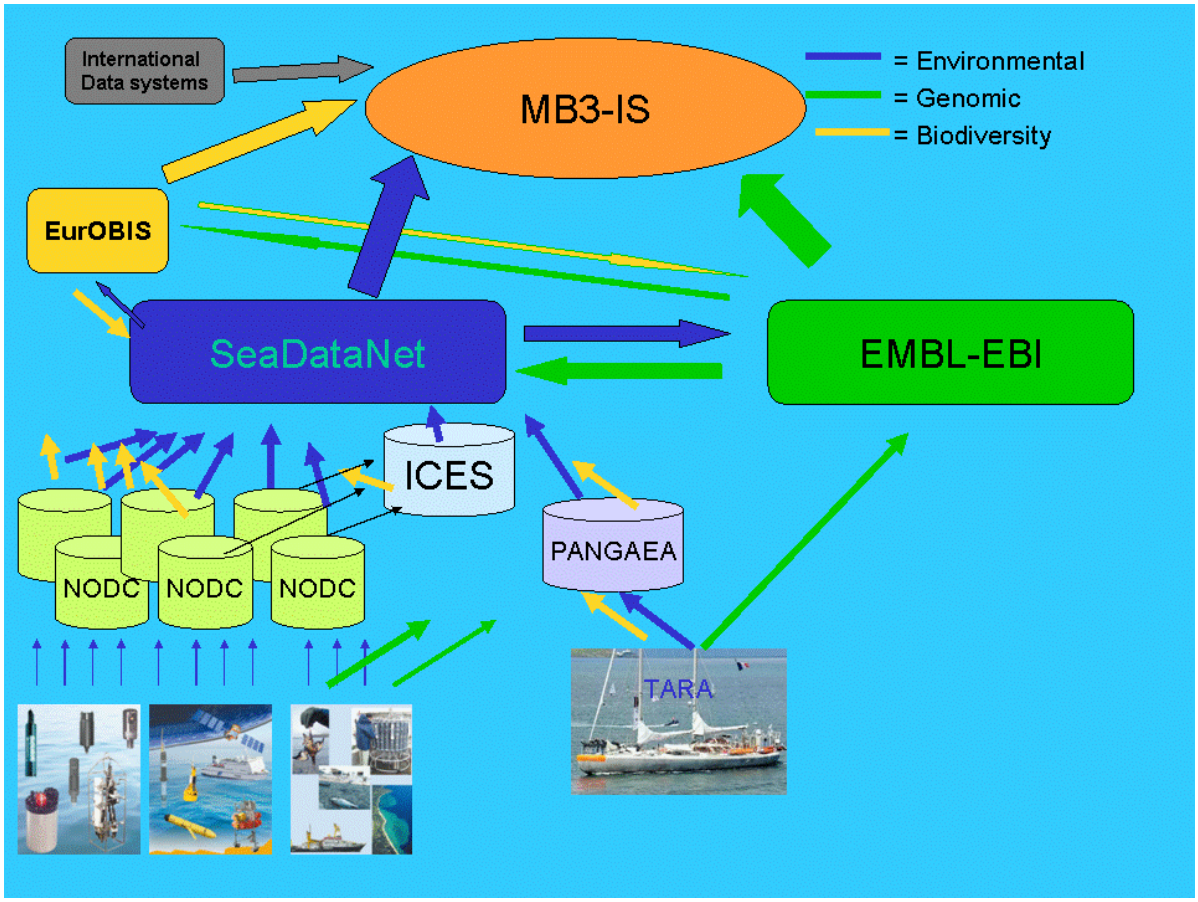

Figure: proposed flow and management of data in Micro B3

Interfacing is planned between the 3 basic infrastructures, SeaDataNet, EurOBIS and ENA, for better service and a complete data provision to internal and external users; within the framework of SeaDataNet II and EMODNet activities are already planned for establishing interoperability between SeaDataNet and EurOBIS; this will be extended in Micro B3 to including ENA in the ring.

Interfacing is also planned from SeaDataNet, EurOBIS, and ENA to MB3-IS to fulfill the specific needs of Micro B3.

Both levels of interfacing will be worked out in more detail in the next chapters.

5.0 Interfacing between SeaDataNet and Micro-B3 Information System (MB3-IS)

The Micro B3 bioinformatics analysis system is developed in WP5. This MB3-IS consists of a software platform to support analysis in three steps:

- Processing for automatic annotation of genomes and metagenomes
- Integrating the outcome of the processing results with marine environmental data such as from SeaDataNet and its underlying databases. Time and space will be the main trigger to combine the microbial data from the laboratory/analyses with the environmental data from the datacenters, but additional keys may be needed
- Visualising the results.

For interoperability of MB3-IS with other data systems 2 main components are of importance:

- The security module: as authentication OpenID has been chosen (Shibboleth is investigated). There will be no additional user management; therefore the solution is quite simple. Consequently from a user only the OpenID url is known, plus maybe the email address. It should be considered whether this is sufficient for accessing the other data infrastructures. SeaDataNet is investigating the use of OpenID on top of its existing CAS login. In addition an OpenID user will have to agree the first time with the SeaDataNet user license.
- MicroB3 catalogue/database: this part of MB3-IS is foreseen to be mainly a catalogue of data services and protocols which are made available using a client to harvest data. MB3-IS will not locally buffer all data. Only in some cases, e.g. for analyses software to run on selected data, the data will be temporarily cached on the MB3-IS server for that specific user.

The discovery and delivery of data sets from SeaDataNet to MB3-IS can be set-up with an increasing service level:

1. Using OGC WMS - WFS protocols to exchange CDI metadata from SeaDataNet to MB3-IS
2. Using OpenID to facilitate the logon from MB3-IS users to SeaDataNet for requesting data sets
3. Developing a machine-to-machine interoperability from MB3-IS to SeaDataNet to provide a more efficient data discovery and delivery service to MB3-IS users

These 3 increasing levels of service will be detailed in the following text.

Ad 1) The SeaDataNet infrastructure has a central portal with the CDI Data Discovery and Access service which gives detailed metadata of the data sets as managed by the distributed data centres (>80 data centres at present). This service also features a data shopping mechanism to support registered users to request access and if granted, to download data sets from the distributed data centres in a common way using the Request Status Manager (RSM) facility at the SeaDataNet portal.

Applying the open standards from the Open Geospatial Consortium (OGC), in particular Web Map Service (WMS) and Web Feature Service (WFS), will allow for sharing maps and features on maps. These services make it possible to share and exchange map layers between internet portals with map viewing services and to interrogate the objects on these maps by clicking on those thereby retrieving feature information. This way SeaDataNet can provide MB3-IS map layers with locations and metadata for observation data sets as included and available in the SeaDataNet CDI Data Discovery and Access Service.

This interoperability solution is already used by SeaDataNet to share CDI metadata with the SIMORC service (SIMORC = System of Industry Metocean data for the Offshore and Research Communities (http://www.simorc.org)). Go to Data Access and activate the extra layer of SeaDataNet Metocean Data. Also activate the i box and use the identification tool to retrieve the CDI metadata. This is illustrated in the following screens of the SIMORC user interface. A direct exchange takes place with the WMS - WFS services as provided by the SeaDataNet CDI portal service. By WMS protocol SIMORC users can activate maps to see the locations of the SeaDataNet data sets. Users can click on sites on the map to retrieve by WFS protocol metadata of that site.

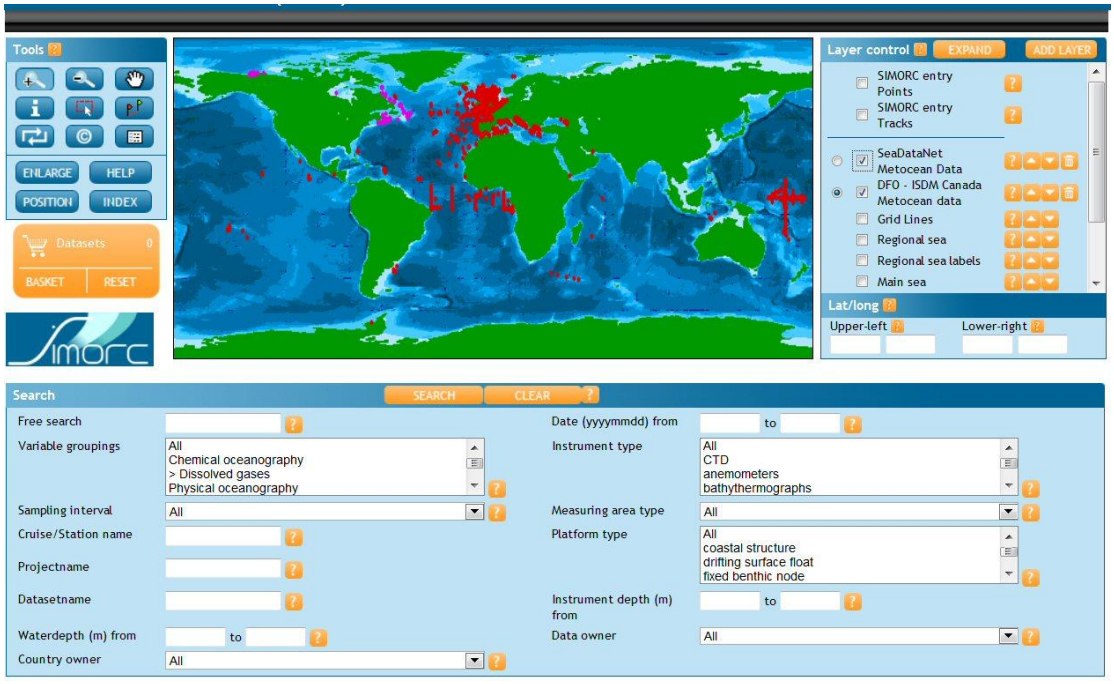

Image: SIMORC User Interface with extra map layer activated for SeaDataNet metocean sites

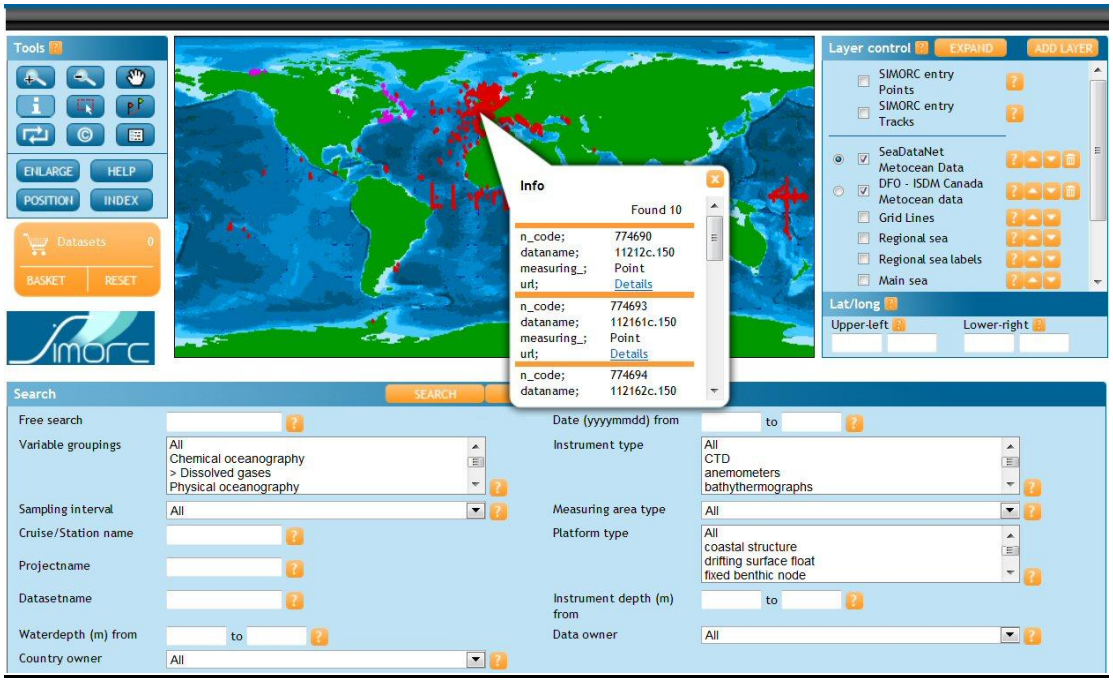

Image: SIMORC User Interface with extra map layer interrogated by WFS to display metadata of SeaDataNet sites

The SeaDataNet metadata as displayed above include also dedicated URLs to additional metadata and data access as provided by the SeaDataNet CDI service. Clicking on the URL will give a link to SeaDataNet for the selected site and data set.

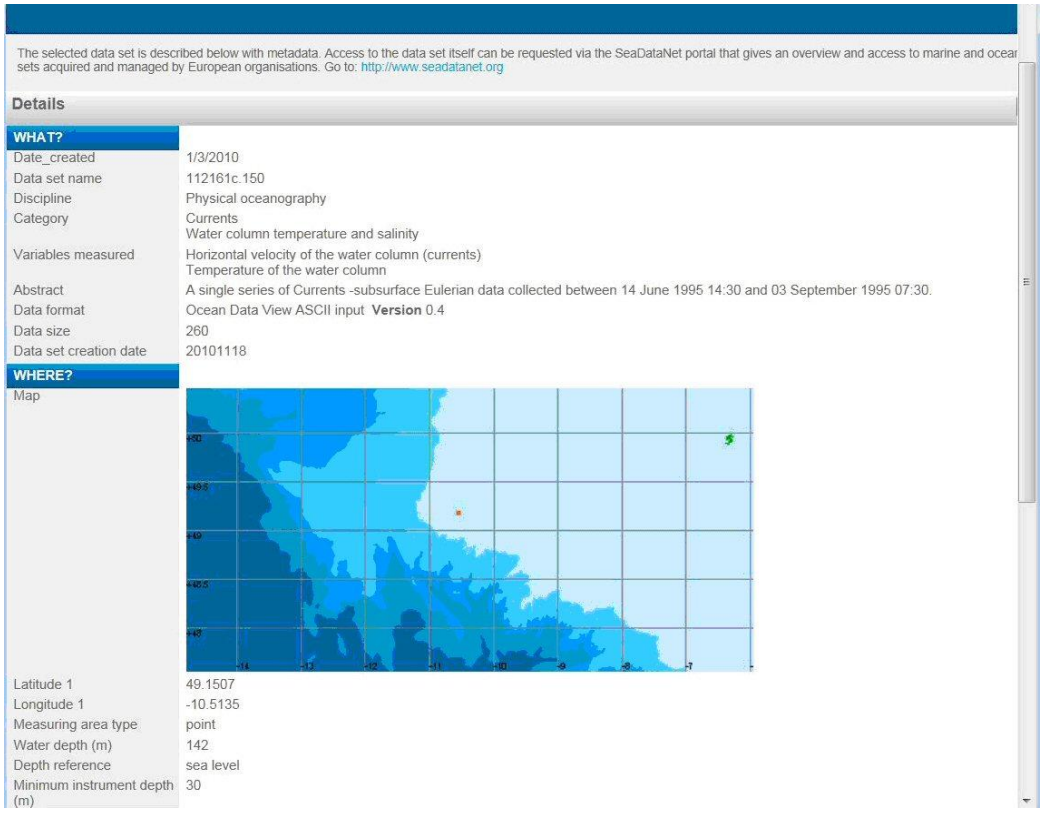

Image: extra metadata as retrieved from SeaDataNet for selected site and data set

In the case of Micro B3 the WMS -WFS coverage of SeaDataNet data sets will not be the full collection, but it will be trimmed / filtered to the specific needs of Micro B3. The required subsetting to specific locations and data types is described in more detail in Deliverable D3.4. The WMS - WFS service for MB3-IS will be set-up in the coming year.

Ad 2) In case the MB3-IS user has discovered interesting data sets and wants to retrieve those, then the user will have to make use of the SeaDataNet shopping procedure. This requires that the MB3-IS user is registered as a SeaDataNet user. MB3-IS will make use of OpenID as authentication protocol. OpenID is an identity provider that can be requested in a standardised way to verify accounts and confirm this. Well known (trusted) systems like Google, Twitter, etc all make use of OpenID and therefore many users already have an OpenID account. It would therefore be very convenient if SeaDataNet can also support OpenID. At present SeaDataNet make use of the CAS system for Authentication, Autorisation and Administration (AAA services) using a central SeaDataNet User Register. Users have to submit their registration requests to SeaDataNet, thereby confirming their acceptance of the SeaDataNet data policy and its license agreement. The shopping mechanism also makes use of roles for users by which a differentiated access process has been effectuated in SeaDataNet. Each CDI metadata record includes a data access restriction tag ranging from ''free open access'' to ''by negotiation''. The combination of user role with the data access restriction drives the access decision matrix to chose between the 3 options: a) having direct automatic access, b) negotiation required, c) access denied. At registration users receive an initial role as public user. Thereafter their role is finetuned by the NODCs of their country, further considering the nature of their organisation.

SeaDataNet is already investigating the use of OpenID on top of its existing CAS login and it is planned to be implemented in the coming 6 months. This will mean that MB3-IS users can use their OpenID also in the SeaDataNet context, providing a Single-Sign-On (SSO) situation, but anyway the MB3-IS users will be required at the first log-on to provide some additional information and to agree with the SeaDataNet License. That way they will become regular SeaDataNet users while using their OpenID credentials.

Ad 3) Having the dedicated Micro B3 WMS - WFS services in place at SeaDataNet, supported with the OpenID Authentication, will make it easier for MB3-IS users to discover and retrieve SeaDataNet data sets. But it still should be considered as a first level interoperability solution, working on a human interface basis.

Therefore there is need for further streamlining the exchange from SeaDataNet to MB3-IS by introducing machine-to-machine interoperability. This will focus on:

- creating a dynamic and internal buffer of data sets at SeaDataNet fitting the Micro B3 purpose

- creating a client service at MB3-IS to prepare, submit and process a shopping list to SeaDataNet of requested data sets

The present CDI system provides a strong foundation because it connects already more than 80 data centres and gives access to a large volume of data sets as managed by these data centres. Large efforts have been undertaken to get to this point and daily efforts are spent to monitor and keep the system fully operational. Therefore for achieving interoperability it is fundamental not to alter the existing CDI system architecture but to develop and adding extra services on top of this foundation. Then the operational system can continue without interruption and a lot of the efforts can be focused on the central CDI system without having to activate all connected data centres.

The CDI shopping system makes use of the central Request Status Manager (RSM) service to communicate with each connected data center for passing the user data requests, preparing the requested data sets, and for giving access to the user by means of downloading from each distributed data center. In case of the SeaDataNet CDI User Interface this RSM process is triggered by human requests and it can take some time to retrieve and deliver all requested data sets to the user. For specific communities like Micro B3 this RSM process can be automated to act as a robot user that maintains an internal buffer of data sets that fit the defined and agreed need of the external community (= so-called Community Data Profile). This profile will specify and determine the specific data needs and also which data providers have agreed to deliver to the specific community for which data access restrictions. Note: the data sets marked for their data access restriction with ''unrestricted access'' and ''SeaDataNet license access'' will have no limits, but it might be that individual data providers agree in addition on a case by case basis to allow access and buffering by the robot to data sets with more restrictive access. Following Deliverable D3.4 such a Data Profile will have to be worked out in more detail for the MB3-IS community as part of the implementation phase. In case of wider access to more restricted data sets, this will have to be underpinned by a written Service Level Agreement between Micro B3 and the specific SeaDataNet data providers agreeing to give wider access to specific data sets for Micro B3 purposes.

As part of SeaDataNet II already a first version of such a robot user process and internal buffer system has been developed, in particular for supporting the production of a Temperature and Salinity Climatology (aggregated data sets product) by MyOcean together with SeaDataNet. For the purpose the robot has been set to discover, request and harvest all freely available T&S data sets from all SeaDataNet data providers into a central buffer database, which has been made available to MyOcean together with the related CDI metadata. This central T&S buffer is further maintained in a dynamic way, following new CDI entries and updates to the central CDI service.

It is planned that for Micro B3 a comparable set-up will be specified and implemented, resulting in a centrally maintained buffer database of data sets that are fit for serving MB3-IS needs and that have been agreed with the associated data providers.

The contents and coverage of the SeaDataNet - Micro B3 buffer can then be communicated to MB3-IS by its metadata following the WMS - WFS solution, that will then have tie-ins to arrange data access. But to make live easier for MB3-IS users it is proposed that also work will be undertaken at MB3-IS to develop a client service that can search on the SeaDataNet offer in the MB3-IS interface and can prepare and follow up shopping lists from MB3-IS to the SeaDataNet Micro B3 buffer.

Such a machine-to-machine interaction is planned with the following specs:

- using SOAP web services
- based on functions of the SeaDataNet RSM (Request Status Manager)
- as security IP recognition is integrated to allow only requests from the trusted MB3-IS portal, plus MB3-IS individual users are to be validated via SDN AAA services (via OpenID support)

The steps in this process are as follows (numbers relate to the image below):

1. Configure: The Micro B3 community makes an agreement with SeaDataNet and its datacenters specifying and agreeing the Micro B3 Community Data Profile. Via the config CMS of the user management a special buffer action is defined (for a certain set of parameters, period and geographic locations)
2. Start order: The robot user is triggered to order and maintain a central buffer database of data according to the set Community Data Profile.
3. Collect: The robot harvests all the requested data sets from the distributed data centers.
4. Check rights: During the process the RSM checks special rights of user in SLA user management, and if necessary can “overrule” normal access restrictions
5. Store: All datasets as well as the CDI metadata are placed as a buffer on the central CDI server
6. Login and create order for subset of defined buffer: A MB3-IS user searches for SeaDataNet data and submits requests via a MB3-IS user interface. This MB3-IS client works on the metadata of the data buffer to which Micro B3 users have access rights. Via the MB3-IS interface the user posts an order. Even though the order can be large the processing is quick because the data is already available. However special attention must be given how to transfer large data sets
7. Download: On a special RSM like user page (AAA login needed), to be developed by SeaDataNet, the MB3-IS user finds his orders. This user page shows:

- Order number and date
- Query details
- Order status
- Download link for zipfile(s)

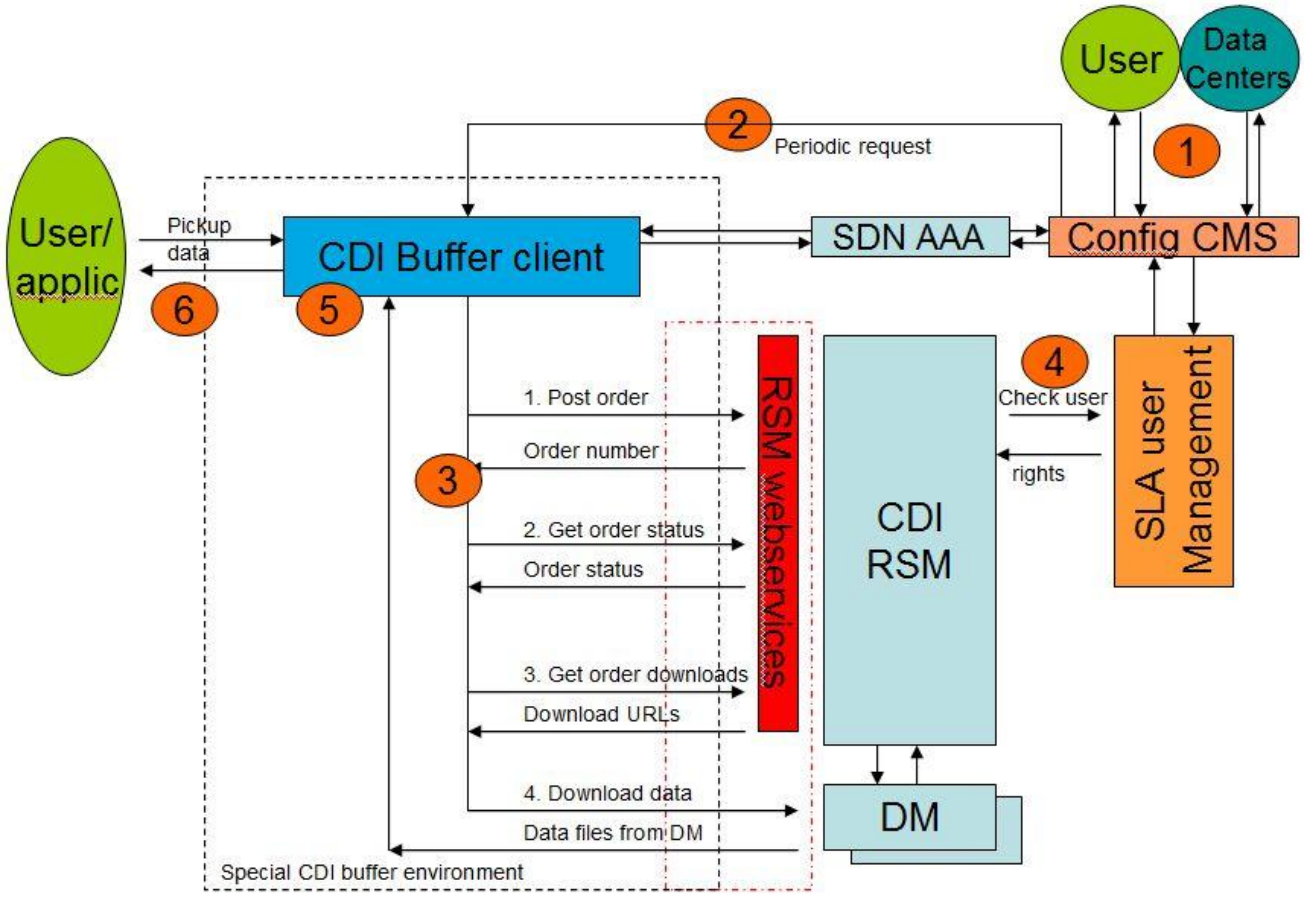

Image: internal buffering at SeaDataNet for Micro B3 profile and followed by exchange with MB3-IS client

This solution will be worked out together with WP5 in the coming year.

6.0 Interfacing between SeaDataNet, EurOBIS and ENA

Interfacing is also planned between the 3 basic infrastructures, SeaDataNet, EurOBIS and ENA, for better service and a complete data provision to their internal and external users. This can be seen as a nice spin-off of Micro B3 that SeaDataNet users will also have direct overview of data sets as managed and provided by EurOBIS and ENA; EurOBIS users of data sets as managed and provided by SeaDataNet and ENA; and ENA users of data sets as managed and provided by SeaDataNet and EurOBIS.

This "horizontal "exchange between the 3 leading European infrastructures can be achieved by the following solutions:

1. Using OGC WMS - WFS protocols to exchange metadata including URLs to further metadata and data
2. Using OpenSearch protocol to facilitate remote queries from one portal to the other portal

Ad 1) applying the open standards from the Open Geospatial Consortium (OGC), in particular Web Map Service (WMS) and Web Feature Service (WFS), as described in the previous chapter. This will make it possible to share and exchange map layers between each of the internet portals with map viewing services and to interrogate the objects on these maps by clicking on those thereby retrieving feature information. The feature info then can include URLs to the associated portal to retrieve further metadata and options to request access and delivery of the data sets.

Ad 2) OpenSearch is another interesting protocol supported by major portals such as Google, Yahoo, Twitter that might be used to facilitate the first level interoperability between the 3 European portals. It provides metadata about the contents along with a set of URL Templates which illustrate the parameters accepted by the service and the variety of output formats in which results can be obtained. The OpenSearch request interface is simple, consisting of a description of a HTTP GET request with a series of optional key-value parameters that can be used to constrain the search:

- Free search
- Geospatial (area or point + radius)
- Temporal (from to)

OpenSearch protocol is rather simple: No use of vocabularies, no special search fields which has a disadvantage when querying a very large collection with specific data like SeaDataNet. It is necessary to split the total metadata in many OpenSearch access points (virtual aggregations).

Format HTTP get:

<Url type="text/html" template="http://example.com/?q={searchTerms?}&pw={startPage?}"/>

Example:

http://www.google.com/?q=question

http://www.google.com/?q={searchTerms}

SeaDataNet already has experience with OpenSearch: an application prototype for OpenSearch has been set up as a result of work in the EU FP7 Genesi-DEC project. An OpenSearch communication layer has been developed on top of CDI database. This prototype is running only for the Geo-Seas subset as yet.

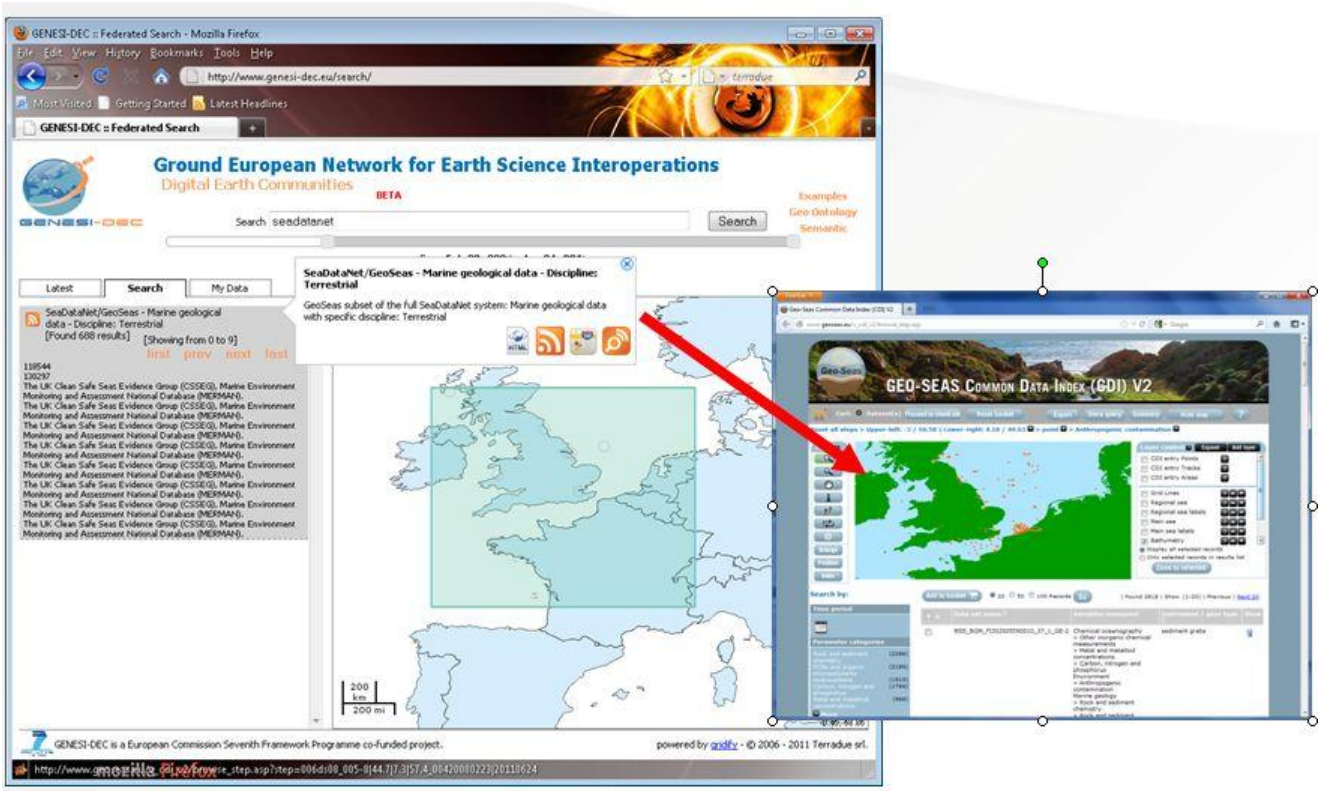

Image: Search result of OpenSearch in Genesi-DEC portal with link to Geo-Seas CDI service for further ordering of related data sets

OpenSearch entry points have been created in RDF and registered to the Genesi-DEC portal (www.genesi-dec.eu). Each RDF file describes an access point of an aggregation of the marine geological data (first filter) and then per discipline (P081).

http://seadatanet.maris2.nl/opensearch/geoseas/ds10/os_description

http://seadatanet.maris2.nl/opensearch/geoseas/B070/os_description

etc.

To provide interoperability between the 3 European portals (SeaDataNet, EurOBIS, ENA) an identical approach can be made for their full collections. The downside of this method is that the virtual series will contain a high number of metadata records (> 100.000). And the OpenSearch protocol is limited in selective search fields. To avoid too many results the best solution is to apply an aggregation per parameter or instrument or data provider when enabling OpenSearch entries.

Example of specific request of NIOZ (NL) data – Per CDI partner per discipline:

http://seadatanet.maris2.nl/opensearch/nioz/ds01/os_result?q=&bbox=&startdate=2005-05-01&stopdate=2010-05-30&pw=1

Following this method for example the SDN CDI directory could be split up in a couple of hundred entry points, which would make the exchange more feasible in practice.

Together with EMBL-ENI (for ENA) and VLIZ (for EurOBIS) further activities will be undertaken in the coming year to set-up the OGC WMS - WFS services and to explore further the benefits and options for adding also OpenSearch services.