Marine
Microbial Biodiversity,
Bioinformatics & Biotechnology

Grant agreement n°287589

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

Deliverable 5.6

Report on interoperability with third party resources

Expected Submission Date: 31.12.2012
Actual submission Date: 20.12.2012

Lead Party for Deliverable: Guy Cochrane, EMBL-EBI

Mail: cochrane@ebi.ac.uk **Tel.: +44 (0)1223 492564**

Dissemination level:	
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)	

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.

GENERALIST SUMMARY

While there is a great deal of new technology to engineer for Micro-B3, the project builds on strong foundations, not least a handful of long-established data resources that span marine science. These data resources include SeaDataNet and its network of National Oceanographic Data Centres (oceanographic data), EurOBIS (macrobiological data) and the European Nucleotide Archive (ENA; molecular sequence data). While these resources all exist to broaden and simplify access to data in their domains, integration of their data across domains requires Micro-B3 researchers to develop a deep understanding of the local organisation and workflows of the domain. In deliverable 5.6, workpackages 3, 4 and 5 have held discussions with this mutual understanding in mind, resolved a 'core' set of data classifiers (those of time and space) that allow cross-domain data sets to be integrated and explored two cross-domain workflow scenarios, those of data submission and data retrieval.

SUMMARY

Micro-B3 workpackage 5, 'Bioinformatics and Data Integration', concerns the design, implementation and presentation of modular software components that will underlie the Micro-B3 Information System (MB3-IS). Part of this work relates to the integration of previously existing oceanographic, biological and molecular data resources which both provide feeds into MB3-IS of existing data and support new data submissions into the system. Oceanographic and biological resources are described in full in D3.1, 'Analysis and selection of oceanographic services for Micro B3' and their services are defined in D3.3, 'Analysis and definition of interoperability of oceanographic services with Micro B3', reports that have been prepared, and should be read, in conjunction with this report. Here, we outline existing submission services and refer to data discovery and retrieval services provided by European Nucleotide Archive (ENA), the molecular data resource, describe a common interoperability 'core' between all Micro-B3 data resources and focus on the two major uses cases: data submissions and data retrieval. This report summarises an ongoing conversation, particularly between those involved in workpackages 3, 4 and 5.

ENA SERVICES

ENA offers a spectrum of data submission, discovery, analysis and retrieval services. Of particular relevance to Micro-B3 are its range of submission services and marine-science-focused programmatic data discovery and retrieval services. These latter have been developed or enhanced under Micro-B3 and are described in full in the report for D5.5, 'Data structures and retrieval services for georeference- and sample variable-oriented ENA data access', so will not be treated further in this report.

ENA submission services are offered that support a range of scales of submission (from single study through ongoing time series to routine aggregation of data from a community and brokering), for a range of data types (from sample and contextual information, raw sequencing reads, assembly information, taxonomic identification tables to functional annotation) according to level of informatics ability/capacity (from fully programmatic webservice-controlled submissions services to interactive web submission applications).

Programmatic submissions

Programmatic submission of data to ENA is available through a coupled secure data drop-box and RESTful webservice system. Users upload data files (such as raw sequence data, read alignments, assemblies and Operational Taxonomic Unit (OTU) tables into their submission account drop-box using FTP, Aspera or, in future, a UDT-based efficient network transfer protocol. ENA offers a command line interface and Java application (the ENA Uploader) that allow users to manage file transfers into their drop-boxes. Once files have been successfully transferred, users call the webservice to issue transactions, that may include requests to validate, accession, load, update, etc. Validation and accession information for any metadata relating to a submitted data set is provided in reports synchronously through the webservice and data file validation, which necessarily takes greater computational time given typical volumes of sequence data files, occurs asynchronously in a process the status of which can be polled regularly through the webservice. Further documentation on ENA programmatic submissions is available from http://www.ebi.ac.uk/ena/about/sra_rest_submissions .

Interactive submissions

The Webin suite of web applications provides interactive submission support for the spectrum of ENA data. The design of these applications focuses on intuitive usability for data submitters. While paths through the applications differ, common themes include modules for registration/login, data management of existing and current submissions, checklist selection (that allows pre-designed checklists of appropriate fields to be captured for a given data submission – for example a MIxS-derived checklist) and spreadsheet-based data entry. Entry into Webin is provided from http://www.ebi.ac.uk/ena/about/submit_and_update.

INTEROPERABILITY 'CORE'

Central to integration of data from independent data resources is a central 'core' of classifiers, or keys, that can be used to relate one otherwise disparate data set to another. Given the breadth of scope of oceanographic, biological and molecular data resources provided into Micro-B3, for example covering information from remote-sensed sea level to molecular function of a given expressed gene, this interoperability 'core' is necessarily limited to a very low level and draws upon descriptors of time and space, universal across these data resources. Table I shows the 'core' classifiers to be used to connect data across resources; each classifier may be supplemented with additional precision and datum information, as appropriate, and with relevant metadata.

TABLE I: INTEROPERABILITY 'CORE'

Classifier	Description
Latitude	Angular distance from equator
Longitude	Angular distance from meridian
Altitude	Distance above or below nominal surface level
Date and time	Date and time of a sampling event
Cruise/station ID	Identifier given to oceanographic cruise or fixed sampling station

DATA SUBMISSIONS

Each of the data resources under analysis (SeaDataNet and its network of National Oceanographic Data Centres[1], EurOBIS and ENA) have established workflows under which they capture data. Recognising the established nature of these workflows, their technical and social complexity, and the numbers of stakeholders depending upon these workflows, it has been clear from very early thinking that existing workflows must be respected and retained, with minor modification and enhancement, where absolutely necessary, to integrate cross-domain submissions. With this in mind, we have developed a view of desired overall data submission workflows (figure I) that we will seek to achieve under Micro-B3. The sources of relevant data are fixed sites (including remote sensors), sampling expeditions and sequencing facilities, to which samples may be sent for molecular sequence analysis from fixed sites and sampling expeditions.

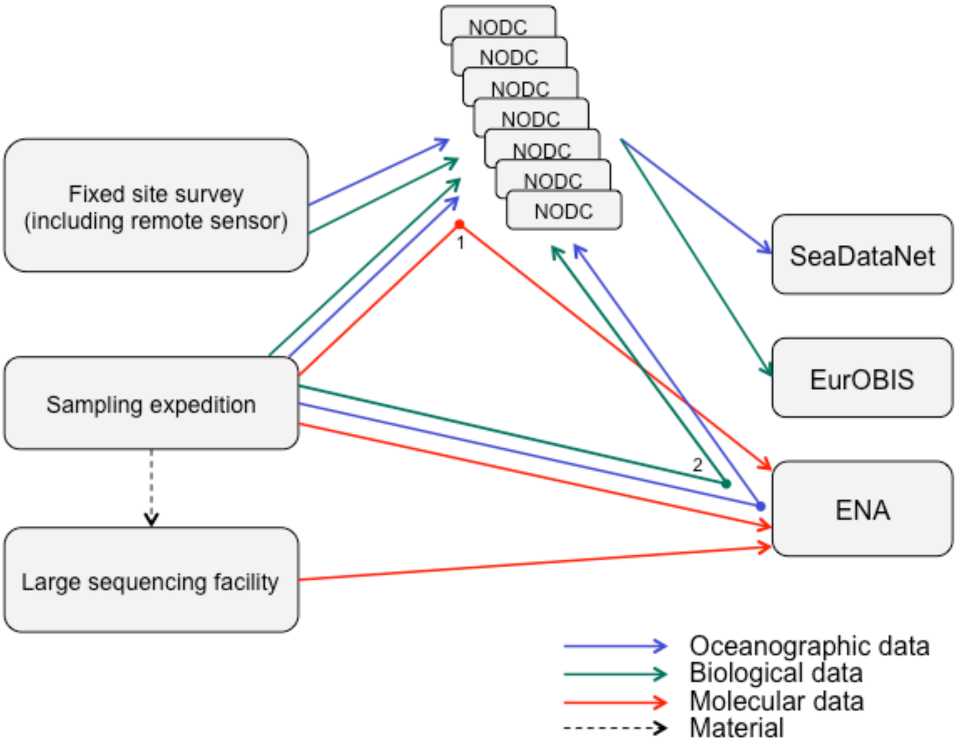
Oceanographic data flow through existing channels to National Oceanographic Data Centres (NODC) and onwards, in summary format, to SeaDataNet. While there is almost no chance that existing or newly established fixed oceanographic observatories will despatch oceanographic data through any routes other than their respective NODCs, there is a chance that new expeditions, in particular those promoted through Ocean Sampling Day and coming from the biological/molecular rather than oceanographic discipline, will approach hosts other than NODCs for their oceanographic data (see figure I, points of inflection 2). In these cases, a page of instructions will be provided by SeaDataNet to ENA that will be used to re-direct (for oceanographic components of data sets) submitters to the appropriate NODCs.

As for oceanographic data, biological data (information on the nature, diversity and abundance of macrobiota in given samples) are routed through existing channels to NODCs. Biogeographic information, however, is despatched from NODCs to EurOBIS rather than to SeaDataNet. Figure I, points of inflection 2 indicates that submitters inappropriately routing this information from sampling expeditions to ENA will be met with a re-direct similar to the above, for which a page will be provided by SeaDataNet for display at appropriate points in ENA submission processes.

[1] NODCs themselves can consist of national centres or national networks of marine institutes, but NODCs seek to provide a national overview of oceanographic and marine data collected and managed by their national institutes.

Molecular data flow through existing channels into the ENA (see details of ENA submission services also above). In figure I, the red arrows depict the flow of sequence data directly from expedition sampling groups and from sequencing facilities. An inappropriate flow, that may be invoked by some submitters, would be of molecular data to NODCs. Figure 1, point of inflection 1 indicates that molecular components of submissions to NODCs will be met with a re-direct. EMBL-EBI will provide an appropriate page to describe the steps an NODC submitter should take re-route this component of the submission directly to ENA.

FIGURE I: DATA SUBMISSION WORKFLOWS BETWEEN DATA RESOURCES.



DATA RETRIEVAL

The provision of an interoperability 'core' of fields allow data integration across the domain-specific data resources. Queries of core classifiers (see table I) are supported uniformly across data resources.

FIGURE II: DATA RETRIEVAL WORKFLOWS ACROSS DATA RESOURCES.

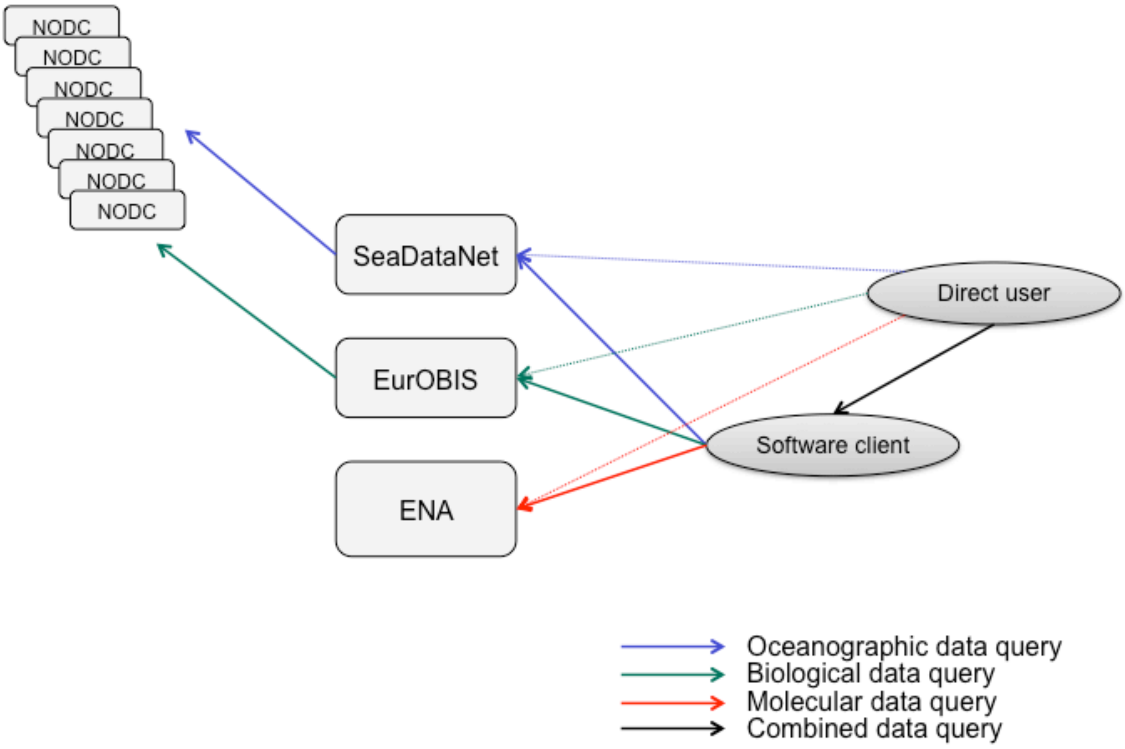


Figure II shows the case where a direct user, or perhaps more ideally a software client, breaks a time-space defined query for data into oceanographic, biological and molecular components. Each of the three data resources provides services to return a report giving information on data holdings against the query. This functionality already exists. For some data resources, such as ENA, direct links to browse and download data are provided, while for others, such as SeaDataNet, for which authentication of incoming users is required to support the licensing framework, these links are not immediately presented. Ultimately, retrieval of deep data, in the case of SeaDataNet and EurOBIS typically requires propagation of data requests down to NODCs. Again, this functionality already exists.