Marine
robial Biodiversity,
ics & Biotechnology

Grant agreement n°287589

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

Deliverable 4.5

Software to support adoption and use of standards when sampling

Version: 1.0
Circulated to: Renzo Kottmann (15.12.2014)
Approved by: WP4 and WP5

Expected Submission Date: 31.12.2014
Actual submission Date: 22.12.2014

Lead Party for Deliverable: EMBL-EBI (Petra ten Hoopen and Guy Cochrane)

Mail: petra@ebi.ac.uk **Tel.: +44 1223 492565**

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.

Summary

Adoption of a data standard can be significantly impacted by the technical support it receives. Even in the case of a well accepted community standard, the proportion of compliant records will be higher if data recording and digitisation are facilitated by an intuitive and user-friendly interface simplifying the reporting process and designed with a deep understanding of standard content. The Micro B3 data-reporting standard is a unique intersection of existing reporting standards of three scientific domains, genomic, oceanographic and biodiversity, developed through a collaboration of several Micro B3 work packages. In order to promote compliance of data records created in data repositories of the three above mentioned domains, work package 4 and 5 developed software tooling supporting marine sampling groups in using the Micro B3 data standard: (1) The European Nucleotide archive has developed and deployed two sample checklists for accurate and compliant description of marine microbial samples, the Micro B3 checklist and Tara Oceans checklist. Although initially designed for specific sampling campaigns these checklists have a broad use and will be fundamental to the description of all marine molecular samples in the public domain. (2) The Micro B3 Information System addressed specific needs of marine laboratories contributing to the Ocean Sampling Day in June 2014 and designed an OSD sample contextual data submission interface that facilitates data reporting and communication of data validation issues to data providers and simplifies data transfer to primary data repositories.

Table of contents

1. Objectives of the Deliverable 4.5

Significant effort has been invested within the Micro B3 project into development of a contextual data reporting system, which minimises the data reporting burden for the sampling laboratories contributing to the Ocean Sampling Day (OSD, http://www.oceansamplingday.org) and at the same time allows creating meaningful data records in data repositories spanning oceanographic, biodiversity and molecular research domains.

The Micro B3 community-developed data reporting standards have been for the first time formulated in the OSD Handbook, Version 1. An updated OSD Handbook, Version 2, was released in June 2014 providing detailed and topical guidelines to the OSD sampling stations on steps to be taken before, during and after the sampling.

Interoperability solutions allowing data archives to share the collected OSD sample contextual information across scientific domains are detailed in Deliverable 4.4. This specifies an interoperability minimum as well as protocols allowing the discovery and delivery of data between components of the Micro B3 infrastructure.

The remaining link in the OSD data flow chain is a software solution supporting sampling groups in recording and digitising the collected sample contextual information, which covers details on the sampling investigation, sample-taking event, sample local environment and sampling protocol.

Deliverable 4.5 describes software tools adapted and developed for support of Micro B3 compliant data reporting. Two main systems serve this purpose and are in details described in the following chapters of this deliverable: (1) the Micro B3 OSD Sample and Metadata Registration portal, a contextual data submission interface designed by the Micro B3 Information System (Micro B3 IS, http://mb3is.megx.net/) and (2) the Micro B3 and Tara Oceans sample checklists developed by the European Nucleotide Archive (ENA, http://www.ebi.ac.uk/ena) for both programmatic and interactive data submissions.

2. Best practice guide for the OSD

The OSD Handbook, Version 1, formulated data reporting requirements for the OSD, guidelines for sampling site registration, sampling permits acquisition, sample archiving and sequencing as well as contextual, environmental and nucleotide data archiving. Due to an intensive development across work packages 2, 3, 4, 5 and 8 an update of the OSD Handbook, Version 2, has been released in June 2014, http://www.microb3.eu/sites/default/files/osd/OSD_Handbook_v2.0.pdf. This version incorporates updated sampling procedures and protocols, OSD data policy and legal permissions framework for sampling, latest workflow for handling OSD samples, sample contextual data, morphology-based biodiversity data and nucleotide sequence data. Changes in these workflows compared to the Version 1 of the Handbook were based on the latest development of the Micro B3 Information System as well as the integration level at involved data archives at the time of the Handbook release.
The OSD Handbook, Version 2, has been successfully used by all participants during the OSD in June 2014 as best practice guidelines for OSD marine stations advising on steps to be taken before, during and after the OSD sampling.

3. Software for OSD sample contextual data submission

3.1 Initial registration

The Ocean Sampling Day, taking place on 21st June 2014, was a simultaneous sampling campaign of one of the largest global-scale networks of marine stations ever made. All sample and contextual data collection was an in kind contribution from the participating marine laboratories.

In order to facilitate, spur on and harmonise transfer of obtained sample contextual data from sampling stations scattered around the world to the relevant European data repositories the Micro B3 Information System has developed a submission portal for initial collection of all OSD sample contextual data, the OSD Sample and Metadata Registration interface.

The OSD Sample and Metadata Registration interface is available from the link: http://mb3is.megx.net/osd-registry/sample-registration [1].
It is a user-friendly interactive form, Figure 1, enabling a user to provide all mandatory and recommended information specified in the OSD Handbook, Version 2. It consists of the following subsections:

Contact details

[1] Source code is available: https://colab.mpi-bremen.de/micro-b3/svn/megx.net/trunk/net.megx.osd.registry/ and https://colab.mpi-bremen.de/micro-b3/svn/megx.net/trunk/net.megx.form-widget/

A submitter contact information section at the top of the form, needs to be completed only once.

Sampling site/event

This section comprises of descriptors characterising the marine sampling campaign, station and OSD sampling event. It corresponds to the content of tables 2b and 2c of the OSD Handbook. All descriptors are expected to have a single occurrence.

Environmental sample

Details of this section are summarised in the table 2d of the OSD Handbook. For each sample, originating from a single sampling protocol, a number of filters can be produced. Each filter sub-sampled from the defined sample can be specified individually.

Environmental data

This section allows reporting on local environment for each sample as specified in the table 1 and 2e of the OSD Handbook.

Information elements addressing morphological identification of organisms found in the sample and methodology of environmental parameter measurement, table 2f and 2g of the OSD Handbook, respectively, are not included in this interface. The oceanographic data centre PANGAEA, http://www.pangaea.de/, will contact the OSD sites and communicate submission of these data directly to PANGAEA using the archive standard operating procedures. It is expected that morphology-based biodiversity information will be available in later stage, and for some marine stations will not be available at all due to diverse research focus and facility base of the OSD marine laboratories.

An **online video tutorial** describing the OSD contextual data submission interface is available at the link:
https://www.youtube.com/watch?v=iSxqjwxMcbM&feature=youtu.be

Centralised collation of the OSD contextual data from all marine laboratories participating in the OSD main event in June 2014 offers several advantages:
(1) It encourages sampling groups to provide data promptly, allowing the OSD management team to proceed with the molecular data analysis,
(2) it allows quality check of submitted data prior to their transfer to relevant primary data archives, the ENA and PANGAEA. The OSD sample contextual data submission portal has incorporated validation rules checking presence of all Micro B3 reporting standard mandatory fields (flagged in the submission form with asterisk) and validating descriptor values against expected data formats,
(3) it enables rapid feedback of reported validation issues to the OSD data provider due to established communication channels between the OSD management team and the OSD participating sampling sites
(4) it facilitates brokering of all OSD contextual data to the above mentioned primary data archives.

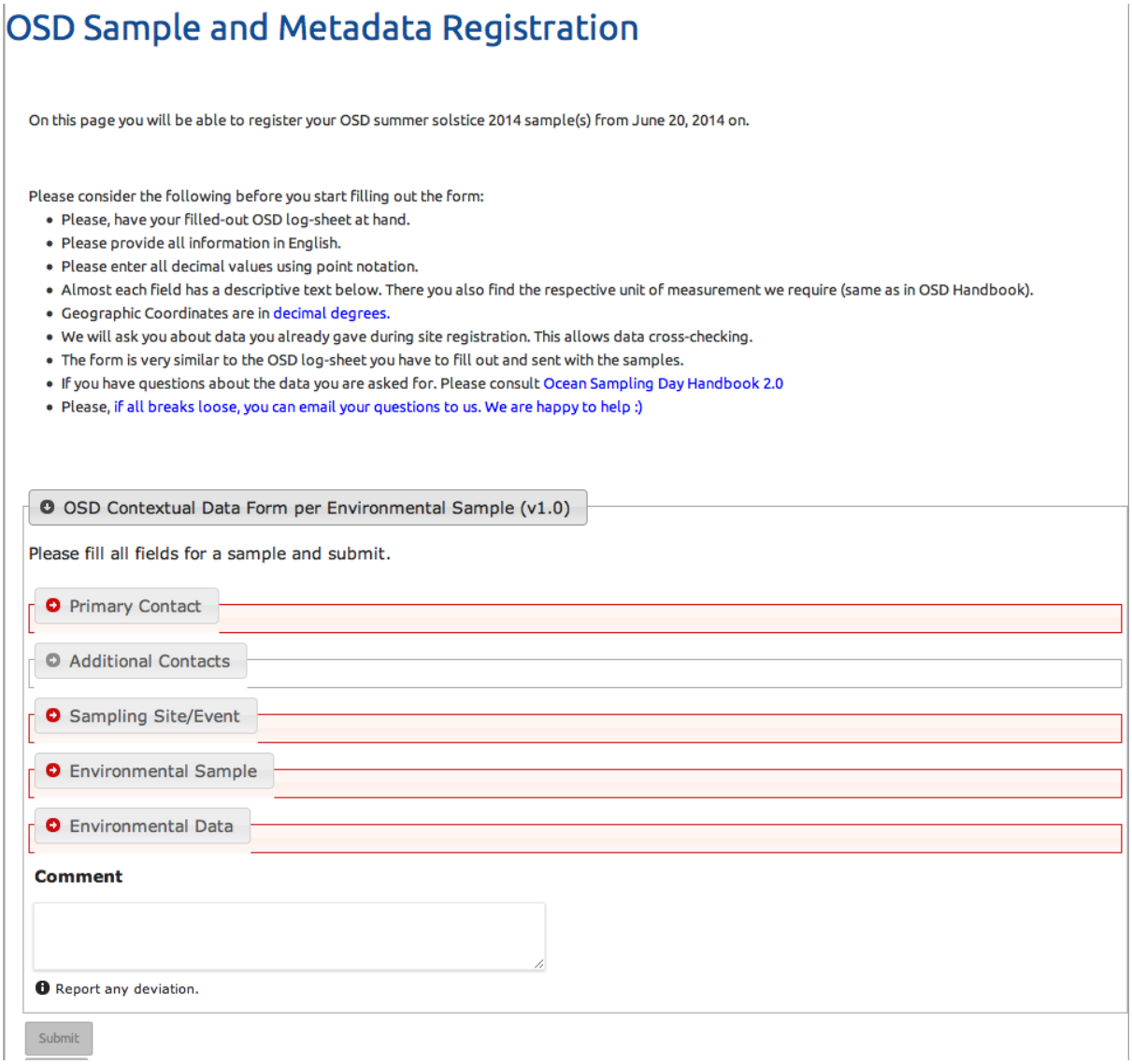

Figure 1: The OSD sample contextual data submission portal developed by the Micro B3 Information System for Micro B3-compliant data reporting.

3.2 Long-term storage and amendments

As mentioned above, the OSD Sample and Metadata Registration interface has been designed for the initial submission of OSD sample contextual data. Long-term storage of the OSD sample contextual data and any amendments are the responsibility of the primary data archives, the PANGAEA and ENA. Both archives will be the long-term guardians and access point of the OSD samples contextual data, with ENA hosting in addition the nucleotide sequence data and PANGAEA the environmental data.

4. Software for marine sample contextual data submission

The Micro B3 data infrastructure aims to support not only the Ocean Sampling Day campaign but also other similar marine enterprises.
This Chapter describes software tooling built within the European Nucleotide Archive to support adoption and use of the Micro B3 reporting standards for marine microbial sampling beyond the Micro B3.

The ENA content is organised around several metadata concepts that aim to provide context to the archived sequence data:

Study – describes focus of the sequencing project
Sample – describes the sequenced biological sample and its surrounding environment
Experiment – describe the DNA/RNA sequencing performed with the sample
Run – describes the raw data files
Analysis – describes data analysis and annotation

All OSD sample contextual data will be transferred from the initial registration data portal described in the Chapter 3 to the ENA, where each OSD sample will be represented as a single sample object receiving a permanent ENA sample accession number.

In order to represent molecular samples accurately and according to community-developed data standards, each ENA sample is described by a sample checklist, which is essentially a list of descriptors tailored to reporting requirements of a community-developed contextual data reporting standard that ENA supports. Currently, a user submitting sequence data to the ENA can choose from over 20 checklists when describing the sequenced samples. These checklists support:

(1) 15 environments defined by the MIxS molecular data standard formulated by the Genomic Standards Consortium (GSC, Yilmaz *et al*., 2011)
(2) the Source feature of the International Nucleotide Sequence Database Collaboration (INSDC, http://www.insdc.org/)
(3) pathogen surveillance samples of the Global Microbial Identifier initiative (GMI, http://www.globalmicrobialidentifier.org/)
(4) prokaryotic pathogen samples
(5) marine samples from the Tara Oceans expeditions (Karsenti *et al*., 2011)
(6) marine samples from the Micro B3/OSD sampling campaign

The two marine sample checklists have been created for the OSD and Tara Oceans enterprises to support Micro B3-compliant contextual data reporting. Both checklists are compliant to the Micro B3 data reporting standard as specified in the OSD Handbook, Version 2.

4.1. The ENA Micro B3 sample checklist

A complete description of the Micro B3 sample checklist can be found at the link: http://www.ebi.ac.uk/ena/submit/microb3-checklist. The checklist is available for both programmatic and interactive data submissions. Figure 2 shows the checklist view in the ENA interactive submission tool WEBIN.

Descriptors of this checklist correspond to the tables 1 and 2 of the OSD Handbook and are divided into the following categories:

1. Marine Sampling
This section contains descriptors of the Sampling Campaign, Site and Platform.

2. Marine Event
This section contains descriptors of the Event Date/Time, Longitude Start, Longitude End, Latittude Start and Latitude End.

3. Marine Sample
This section contains descriptors of the Depth and Protocol Label.

4. Marine Environment
This section contains descriptors of the Marine Region, Environment (Biome), Environment (Feature), Environment (Material), Temperature and Salinity.

5. Environmental Conditions
This section contains environmental descriptors, such as alkalinity, conductivity or fluorescence.

6. Concentration Measurements
This section contains environmental parameter descriptors, such as concentration of nutrients, bacterial production and respiration.

7. Sample Collection Terms
This section contains environmental descriptors, such as biomass, organism count or sample density.

8. Organism Characteristics
This section contains environmental descriptors, such as trophic level or known pathogenicity.

While the sections 1-4 allow provide the Micro B3 mandatory reporting elements as described in the table 2a, the sections 5-8 allow to define the recommended descriptors mentioned in the tables 1 and 2b – 2e of the OSD Handbook.

Although development of the ENA Micro B3 sample checklist has been initiated by the primary sampling campaign of the Micro B3 project – the OSD, this marine checklist can describe marine molecular samples of other marine campaigns. Indeed, the ENA already

recorded an interest of other marine initiatives to use this checklist for registration of their marine molecular samples at the ENA.

In order to facilitate the sample registration process at the ENA, a step-by-step tutorial has been created explaining how to use the WEBIN submission system and the Micro B3 checklist for marine molecular sample submission.

The **online tutorial** can be found at the link:
https://www.youtube.com/watch?v=ZKF1DVs6Lbo&feature=youtu.be

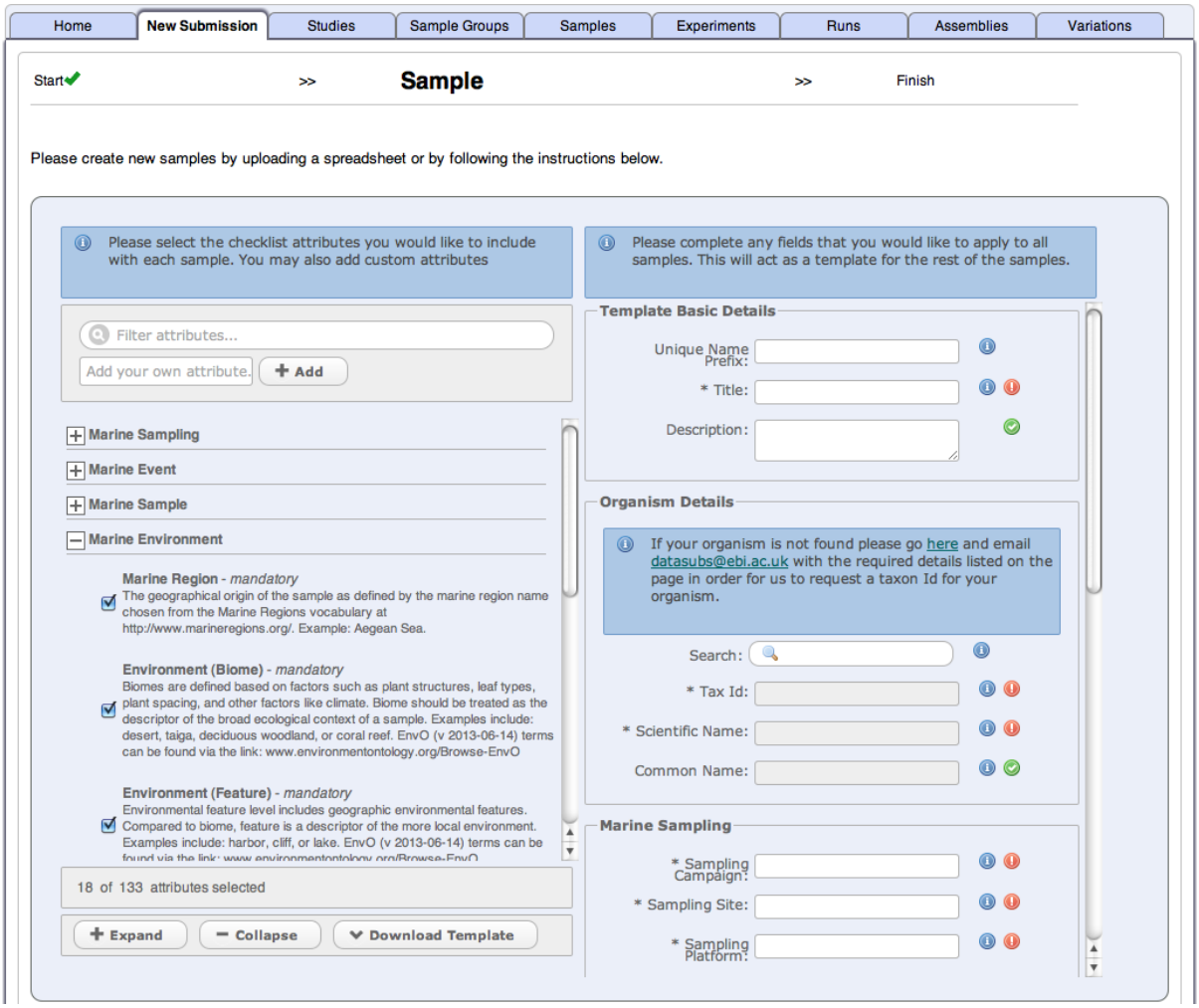

Figure 2: The ENA Micro B3 checklist in the interactive WEBIN submission system developed by the European Nucleotide Archive for contextual data submission of marine molecular samples.

4.2. The ENA Tara Oceans sample checklist

A complete description of the Tara Oceans sample checklist can be found at the link: http://www.ebi.ac.uk/ena/submit/tara-oceans-checklist. The checklist is available for both programmatic and interactive data submissions. Figure 3 shows the checklist view in the ENA interactive submission tool WEBIN.

A high level of similarity of this checklist with the ENA Micro B3 checklist is not surprising since both checklist are compliant to the same Micro B3 data reporting standard (summarized in the OSD Handbook, Version 2).
The Tara Oceans checklist has been developed in collaboration with PANGAEA, the data management centre of all Tara Oceans data and the provider of all contextual data to the ENA. Differences reflect specific needs of the Tara Oceans consortium.

The ENA Tara Oceans sample checklist contains the following categories of descriptors:

1. Marine Sampling
This section contains descriptors of the Sampling Campaign, Site and Platform.

2. Marine Event
This section contains descriptors of the Event Date/Time Start, Event Date/Time End, Longitude Start, Longitude End, Latitude Start, Latitude End and Event Label.

3. Marine Sample
This section contains descriptors of the Depth and Protocol Label, Sample Collection Device, Size Fraction Lower Threshold, Size Fraction Upper Threshold, Sample Status and Last Update Date.

4. Marine Environment
This section contains descriptors of the Marine Region, Environment (Biome), Environment (Feature), Environment (Material), Temperature and Salinity Sensor, Oxygen Sensor, Nitrate Sensor and Chlorophyll Sensor.

5. Reference to PANGAEA
This section describes references to the Tara Oceans Sample Registry at the PANGAEA

Although designed for accurate description of the Tara Oceans samples, the checklist is publicly available and well suited for description of molecular samples originating from the marine pelagic zone.

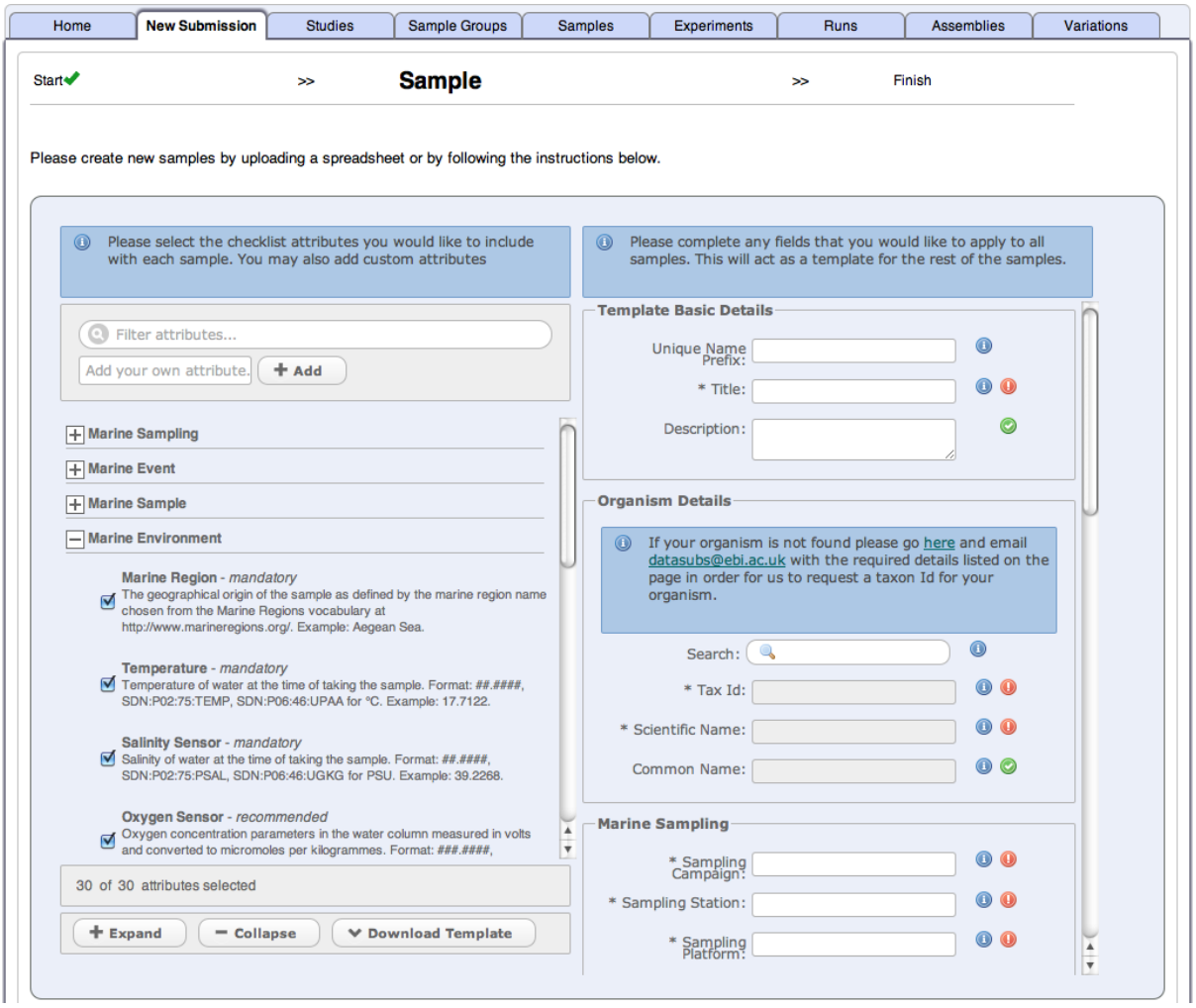

Figure 3: The ENA Tara Oceans checklist in the WEBIN submission system developed by the European Nucleotide Archive for contextual data submission of marine molecular samples of the Tara Oceans Expeditions.

Reference list

Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L, Gilbert JA, Karsch-Mizrachi I, Johnston A, Cochrane G *et al.* (2011) Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nature Biotechnology 29: 415–420.

Karsenti E, Acinas SG, Bork P, Bowler C, De Vargas C, Raes J, Sullivan M, Arendt D, Benzoni F, Claverie JM *et al.* (2011) A holistic approach to marine eco-systems biology. PLoS Biology 9(10): e1001177.