Marine

Microbial Biodiversity,

Bioinformatics & Biotechnology

Grant agreement n°287589

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

Deliverable 4.7

Software to simplify and accelerate data flow from sampling groups towards ENA

Version: 1.0
Circulated to: Guy Cochrane (EMBL-EBI), Petra ten Hoopen (EMBL-EBI), Frank Oliver Glöckner (Jacobs Uni)
Approved by: Frank Oliver Glöckner (15/12/2015)

Expected Submission Date: 31/12/2015
Actual submission Date: 16/12/2015

Lead Party for Deliverable: EMBL-EBI
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 has received funding from the European Union’s Seventh Framework Programme for research, technological development and demonstration under grant agreement no 287589 (Joint Call OCEAN.2011-2: Marine microbial diversity – new insights into marine ecosystems functioning and its biotechnological potential).
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

An intuitive and user-friendly interface simplifying publication of data as well as mature data discovery services are essential for the support and encouragement of scientists in publishing data associated with their research. The Micro B3 project has enabled EMBL-EBI's European Nucleotide Archive (ENA) to develop and deploy software that simplifies reporting of marine nucleotide data and associated information to the archive and software that improves discoverability of the archived data. While marine microbial research was the direct use case for this development, many of the enhancements will serve the needs of broader environmental genomics, the marine metagenomics community and beyond.

Two marine campaigns supported throughout the Micro B3 project, Tara Oceans and Ocean Sampling Day, now have available wealths of rich and valuable data that must reach the public domain for downstream analysis and use. Through ENA, we have facilitated the reporting of the marine shotgun and amplicon raw read data, genomic contigs and pan-genome gene calls as well as associated metadata. The team has also been heavily involved in the presentation of the data, and dissemination of the work, to the scientific community and the broader public, through data analysis workshops and media communication, respectively.

Table of contents

1. Objectives of the Deliverable 4.7

Standards and technologies supporting the reporting of marine data into structured data repositories for long-term preservation with maximum discoverability and re-usability are a key legacy output of Micro B3. Building on Deliverable 4.5, in which we advanced data reporting software and marine standards, in Deliverable 4.7, we further simplify data reporting workflows, build sustainability into the management of standards and enhance the visibility of marine data.

The main objective of Deliverable 4.7 is to develop software to support simplified and more effective reporting of marine molecular data and associated information to the European Nucleotide Archive (ENA; http://www.ebi.ac.uk/ena) as well as improved discovery of richly contextualized sequence data from marine projects such as Tara Oceans or Ocean Sampling Day.

Here we summarise work done on both the data submission and retrieval side of the archive:

- We have developed a sample checklist-editing environment which enables curators to design new, and maintain existing, checklists of sample-related information in more consistent, efficient and sustainable ways.
- We have made advances to the EMBL-EBI submission system for those users reporting functional annotation on their sequence data, including marine data.
- We have added new functionalities for better discovery of environmental sequence data at ENA, such as localisation of a sample geographic provenance on a map or a new Environmental domain in the ENA Advanced Search service, supporting search on additional fields relevant to environmental data.
- We have facilitated deposition of results from sequencing efforts of Ocean Sampling Day 2014 and Tara Oceans to the ENA.

We also report under Deliverable 4.7 extensive outreach activities related to communication of Tara Oceans and Ocean Sampling Day 2014 to the broader research community, media and public.

2. ENA sample checklists

2.1 Web-based sample checklist curation and validation environment

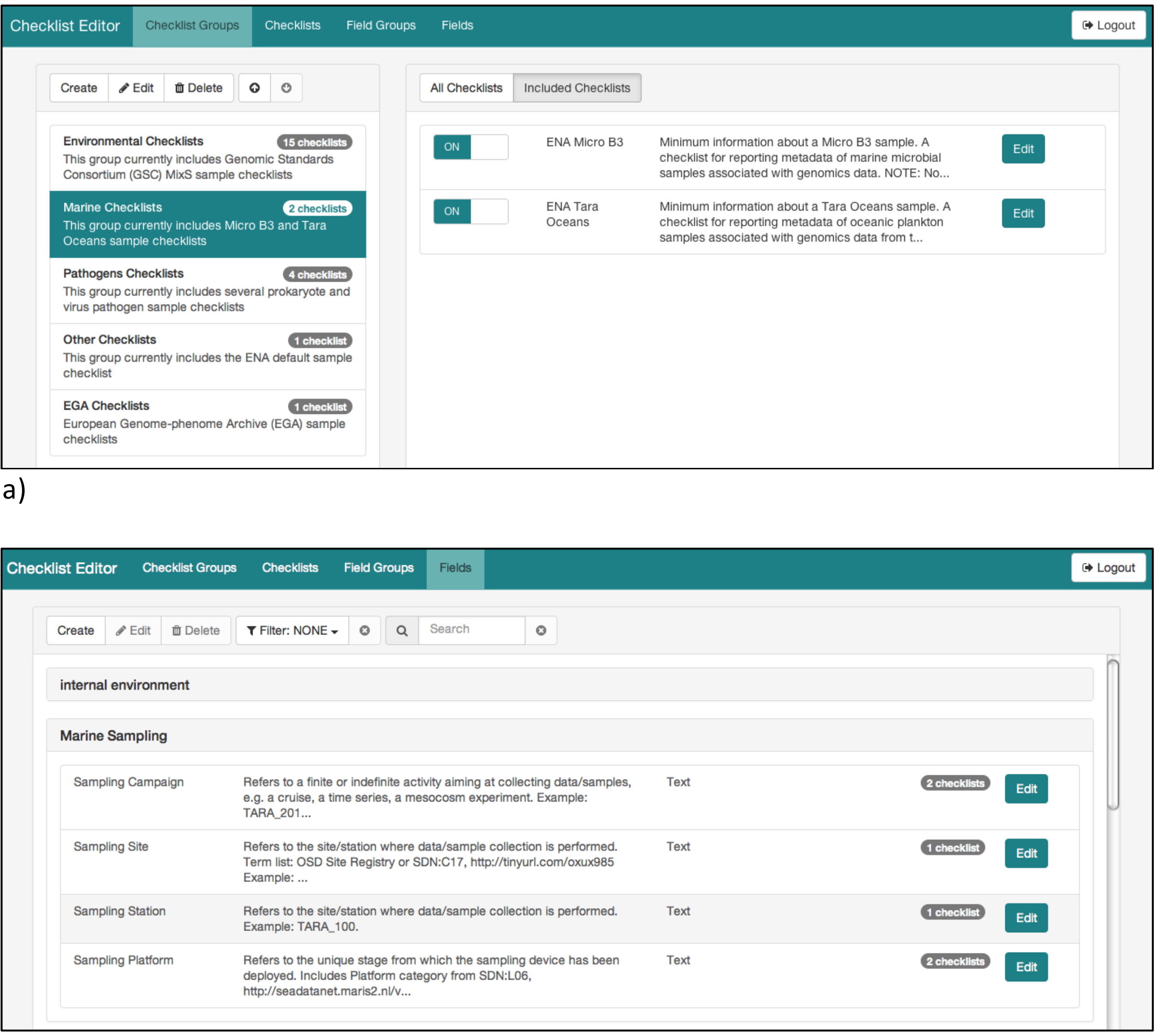
Contextual data capture is one of the essential aspects of the ENA operation. Diverse expert communities collaborate with the ENA team on the development of minimum information recommendations, yielding an ever-growing list of ENA sample 'checklists'. While these bring greater consistency and increased value to the data that they support, maintaining integrity within checklist and consistency between them is challenging for those responsible for the data reporting, validation and presentation services. EMBL-EBI has therefore invested considerable energy into the development of a web-based sample checklist curation and validation environment, Figure 1, that enables the ENA content team to manage the ensemble of checklists efficiently and consistently and has led to the development of new checklist-related functionalities.

The fundamental conceptual unit in this editor is a checklist field (descriptor), which can have various formats (text, text area, regular expression, controlled vocabulary, ISO8601 compliant date & time or taxonomy). Each checklist field can (1) belong to one field group, (2) be associated with 1 to N checklists, (3) be flagged as mandatory/recommended/optional, (4) be present in the checklist single or multiple times and, importantly, (5) be propagated across all relevant field groups and checklist groups, allowing consistent evolution of each field.

Newly added functionalities include (1) a RESTful system supporting synonyms look-up for checklist fields and controlled values, (2) a RESTful system that increases tolerance to punctuation variation in checklist fields and controlled values, where automated fix of the field or value is applied if match is found to an established field or value, (3) adaptations to the WEBIN data submission system to support checklist groups view, taxonomy and date & time field formats, (4) adaptations to tabular submissions workflows in WEBIN to support sample checklists, such as improved matching of similar field or controlled values.

A deployment protocol has been developed that, following checklist curation activity, transforms the checklist editor content into XML-formatted files to be used by the WEBIN application and the ENA Browser (details in the chapter 2.2).

Since changes made within the sample checklist curation and validation environment can have a serious impact across all checklists we have implemented two content reversion mechanisms that can rectify possible human errors: (1) a rapid re-construction of the editor schema from a previously deployed stage and (2) rapid reversion to previously deployed XML files.



Environmental Checklists — 15 checklists
This group currently includes Genomic Standards Consortium (GSC) MixS sample checklists

Marine Checklists — 2 checklists
This group currently includes Micro B3 and Tara Oceans sample checklists

Pathogens Checklists — 4 checklists
This group currently includes several prokaryote and virus pathogen sample checklists

Other Checklists — 1 checklist
This group currently includes the ENA default sample checklist

EGA Checklists — 1 checklist
European Genome-phenome Archive (EGA) sample checklists

ON	ENA Micro B3	Minimum information about a Micro B3 sample. A checklist for reporting metadata of marine microbial samples associated with genomics data. NOTE: No...	Edit
ON	ENA Tara Oceans	Minimum information about a Tara Oceans sample. A checklist for reporting metadata of oceanic plankton samples associated with genomics data from t...	Edit

(a)

internal environment

Marine Sampling

Sampling Campaign	Refers to a finite or indefinite activity aiming at collecting data/samples, e.g. a cruise, a time series, a mesocosm experiment. Example: TARA_201...	Text	2 checklists	Edit
Sampling Site	Refers to the site/station where data/sample collection is performed. Term list: OSD Site Registry or SDN:C17, http://tinyurl.com/oxux985 Example: ...	Text	1 checklist	Edit
Sampling Station	Refers to the site/station where data/sample collection is performed. Example: TARA_100.	Text	1 checklist	Edit
Sampling Platform	Refers to the unique stage from which the sampling device has been deployed. Includes Platform category from SDN:L06, http://seadatanet.maris2.nl/v...	Text	2 checklists	Edit

(b)

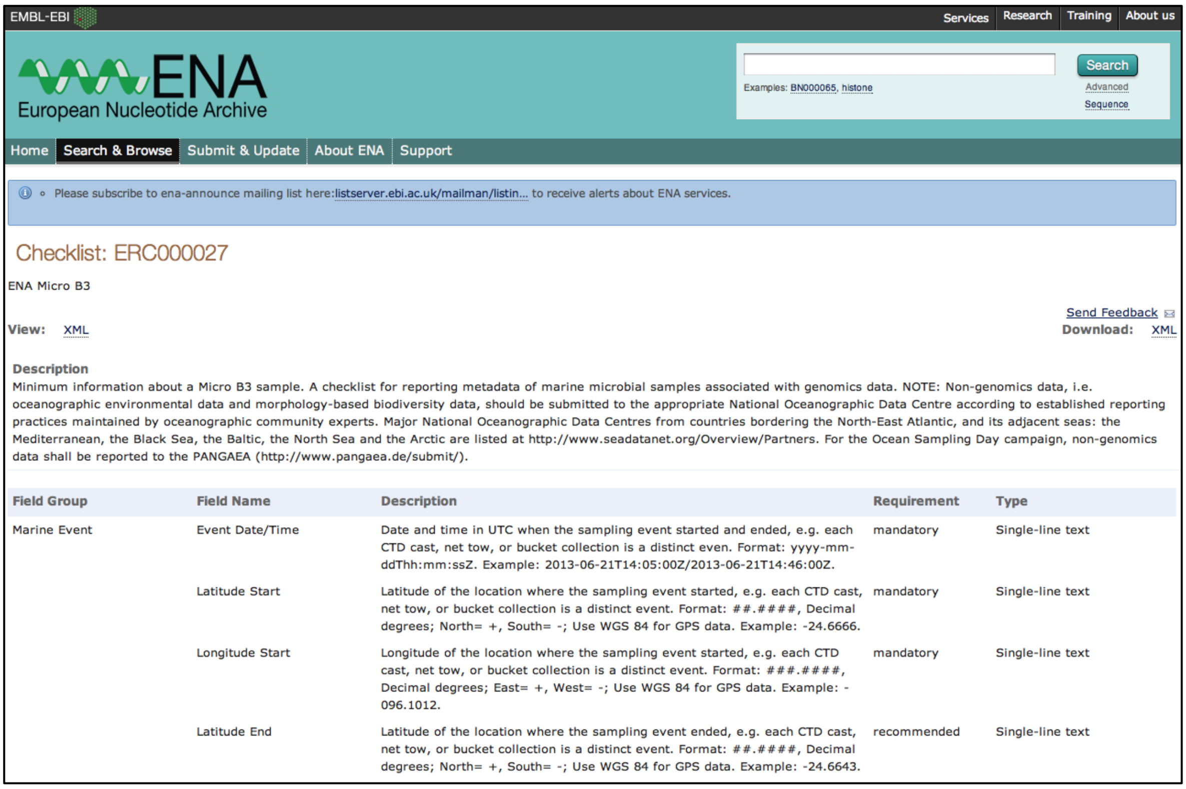
Figure 1: Sample checklist curation and validation environment.

(a) the *Checklist Group* tab: the left-hand panel lists available checklist groups and the right-hand panel lists all checklists within the checklist group;

(b) the *Field* tab: an example of one field group, the Marine Sampling, is shown here with all fields (sample attributes) within the field group.

2.2. Rendering of sample checklists in the ENA browser

In order to disseminate reporting standards and to provide transparency for those using validation tools based on these standards, services to present documentation on checklists have been put in place. The ENA sample checklist content is, upon deployment into submission systems, displayed in human readable form within the ENA Browser, Figure 2. This allows submitters to always refer to the latest version of the sample checklists.



Please subscribe to ena-announce mailing list here:listserver.ebi.ac.uk/mailman/listin... to receive alerts about ENA services.

Checklist: ERC000027

ENA Micro B3

Send Feedback

View: XML

Download: XML

Description

Minimum information about a Micro B3 sample. A checklist for reporting metadata of marine microbial samples associated with genomics data. NOTE: Non-genomics data, i.e. oceanographic environmental data and morphology-based biodiversity data, should be submitted to the appropriate National Oceanographic Data Centre according to established reporting practices maintained by oceanographic community experts. Major National Oceanographic Data Centres from countries bordering the North-East Atlantic, and its adjacent seas: the Mediterranean, the Black Sea, the Baltic, the North Sea and the Arctic are listed at http://www.seadatanet.org/Overview/Partners. For the Ocean Sampling Day campaign, non-genomics data shall be reported to the PANGAEA (http://www.pangaea.de/submit/).

Field Group	Field Name	Description	Requirement	Type
Marine Event	Event Date/Time	Date and time in UTC when the sampling event started and ended, e.g. each CTD cast, net tow, or bucket collection is a distinct even. Format: yyyy-mm-ddThh:mm:ssZ. Example: 2013-06-21T14:05:00Z/2013-06-21T14:46:00Z.	mandatory	Single-line text
	Latitude Start	Latitude of the location where the sampling event started, e.g. each CTD cast, net tow, or bucket collection is a distinct event. Format: ##.####, Decimal degrees; North= +, South= -; Use WGS 84 for GPS data. Example: -24.6666.	mandatory	Single-line text
	Longitude Start	Longitude of the location where the sampling event started, e.g. each CTD cast, net tow, or bucket collection is a distinct event. Format: ###.####, Decimal degrees; East= +, West= -; Use WGS 84 for GPS data. Example: -096.1012.	mandatory	Single-line text
	Latitude End	Latitude of the location where the sampling event ended, e.g. each CTD cast, net tow, or bucket collection is a distinct event. Format: ##.####, Decimal degrees; North= +, South= -; Use WGS 84 for GPS data. Example: -24.6643.	recommended	Single-line text

Figure 2: The ENA Micro B3 checklist for contextual data submission of marine molecular samples available for view in the ENA Browser (http://www.ebi.ac.uk/ena/data/view/ERC000027).

3. Programmatic interface for ENA non-genome scale sequence and annotation submission

ENA accepts nucleotide sequence data ranging from raw reads to highly derived assembled and annotated sequences. While all data types are submitted using the WEBIN framework, very distinct workflows can exist for different data types with different levels of support across the programmatic and interactive interfaces offered by the system. Study registration, sample reporting and raw data file uploads are effected in one submission workflow, and enjoy support in interactive and programmatic WEBIN interfaces. Assembled/clustered non-genome scale sequences, however, have until now only been supported in an interactive interface. Impacted by this are metabarcoding studies from users wishing to take advantage of programmatic interfaces, such as those providing marine prokaryote diversity data using 16S marker amplicon libraries; these users have so far not had access to any programmatic interface.

To address this issue, we have extended the WEBIN programmatic submission interface to include support for non-genome scale sequence submissions. Repurposing technology form the existing assembly submission pipeline, a workflow has been deployed in which the user (1) defines a sequence class (such as rRNA), (2) associates (optionally) sequences with an existing sample record (3) refers to a sequence data file(s) and a tabular annotation file(s) in the user upload area, and, having routed these data through validation, feeds conventional flatfile sequence records into ENA. This new functionality now allows programmatic submission, in a single operation, of data sets that combine raw data and derived sequences, such as those of marine metabarcoding study submitters.

The sequence template expansion triggered additional improvements of non-genome scale sequence validation, such as a validation check rejecting the qualifier /number on UTRs feature, a check ensuring that translation table 11 is applied on plastids, a check that removes the qualifier /map on STS feature or validation of common names in the /host qualifier of the source feature.

The unified programmatic foundation that we now have in place, will allow in future for the development of a single interactive WEBIN workflow for data sets that span raw and derived data that removes redundancy (in sample reporting) and simplifies the user experience.

4. Advances for environmental sequence data discovery

Long-term storage of the Micro B3-produced data is the responsibility of the relevant primary data archives, PANGAEA and ENA. Both archives are the long-term guardians and access point for OSD and Tara Oceans data, with ENA hosting the nucleotide sequence data and PANGAEA the environmental data. This responsibility requires the archives to continuously invest into improvement of data discovery services.

ENA has made significant enhancements in its discovery and presentation layer for environmental sequence data:

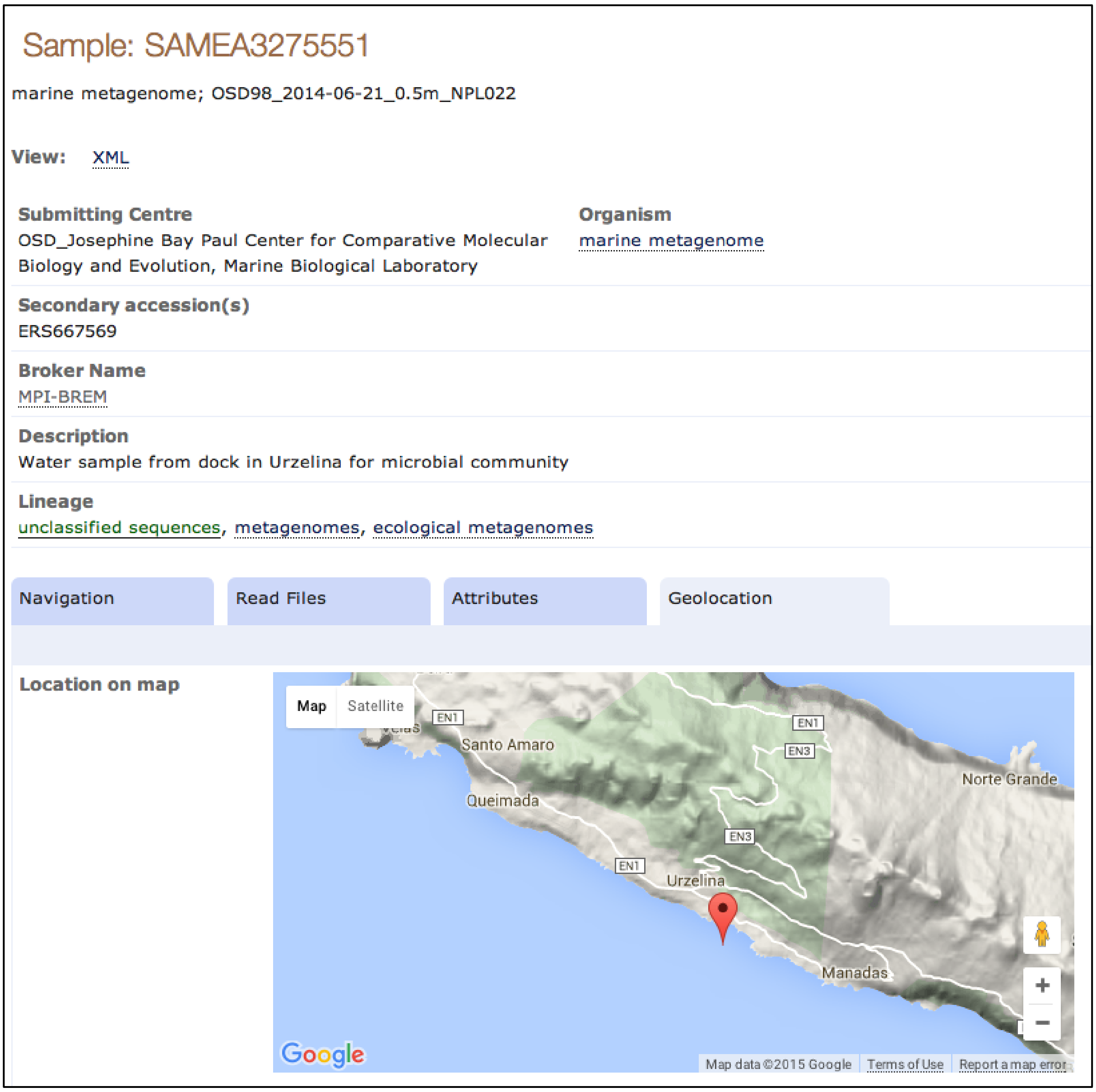
(1) A new E*nvironmental domain* has been added to the ENA Advanced Search service improving search on environmental sequence data using an extended list of indexed contextual data attributes relevant specifically to the environmental data. For instance, the attributes ‘Marine Region’, ‘Sampling Station’ and ‘Sampling Site’ are now individually searchable and the geographic location search allows users to edit the southwest and northeast points.

(2) New browser functionality has been introduced in the Sample domain enabling users to locate a sample geographic provenance on a map, Figure 3. This *geolocation* tab displays a zoomable map with the geographical coordinates that were reported during the sample submission.

(3) The *parent project ID* can now be used for searches in the Read and Analysis domains.

(4) Data in the Read, Analysis and Study domains can now be searched for by the *broker name* and the Run view of the ENA browser specifies the broker name and provides a link to the particular institute, which brokered the data submission, Figure 3.

(5) The controlled vocabulary of checklists in the Sample domain Advanced Search displays names of the checklists but the search query is built using the checklist ID. This allows a user-friendly identification of a sample checklist by its name and unique specification of a checklist in the query builder using its checklist ID that is not meaningful to the user.



Sample: SAMEA3275551

marine metagenome; OSD98_2014-06-21_0.5m_NPL022

View: XML

Submitting Centre
OSD_Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory

Organism
marine metagenome

Secondary accession(s)
ERS667569

Broker Name
MPI-BREM

Description
Water sample from dock in Urzelina for microbial community

Lineage
unclassified sequences, metagenomes, ecological metagenomes

Figure 3: The location of one OSD sample on a map in the *Geolocation* tab of the ENA sample record (http://www.ebi.ac.uk/ena/data/view/ERS667569).

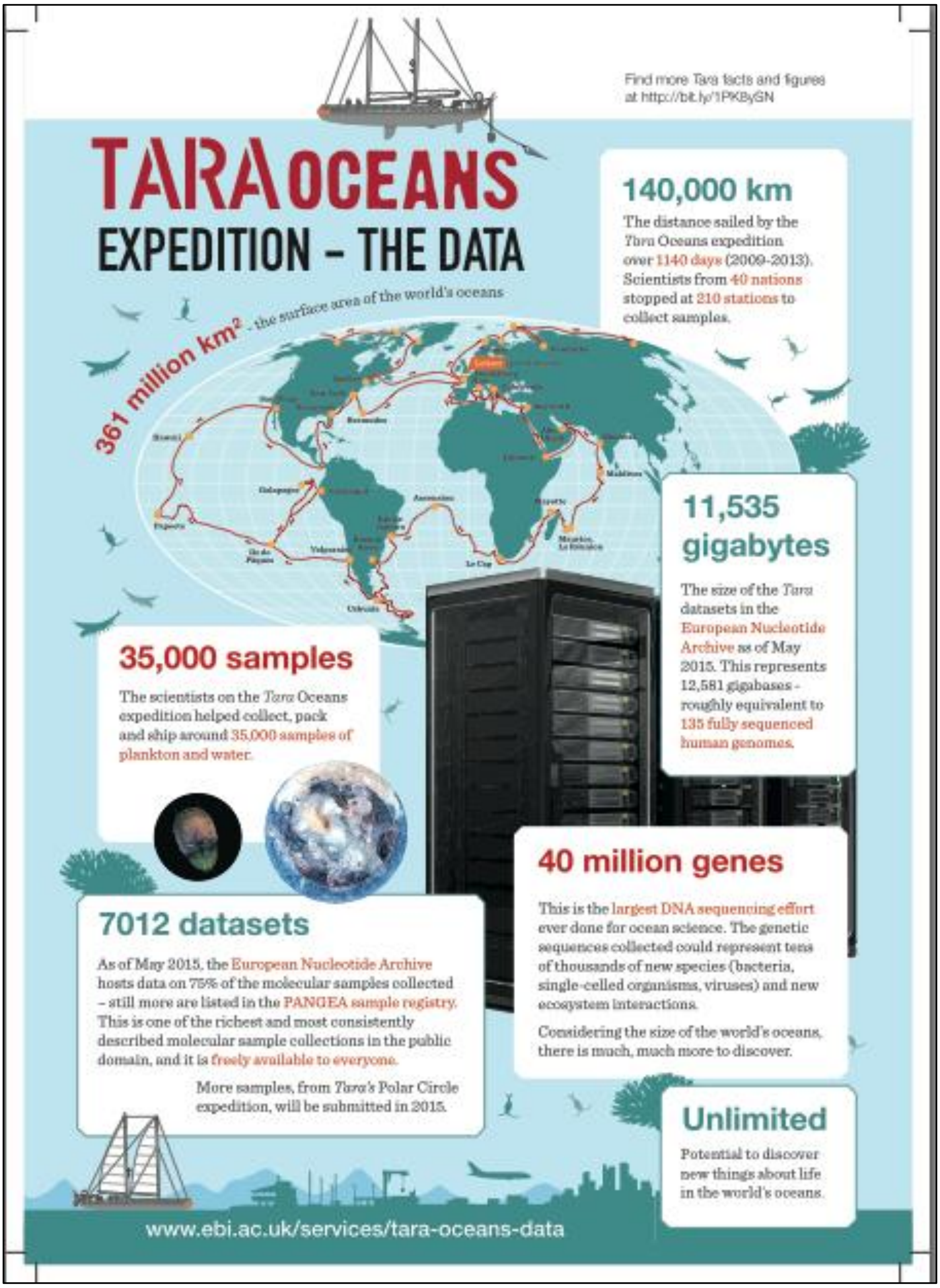
5. Support for Tara Oceans data presentation

The Tara Oceans project has been this year in the spotlight due to a special edition of five Tara Oceans scientific articles published in Science in May 2015, (Brum *et al.* 2015, De Vargas *et al.* 2015, Lima-Mendez *et al.* 2015, Sunagawa *et al.* 2015, Villar *et al.* 2015). With ENA as the archive of molecular data linked to these scientific articles, the team provided intensive support for timely presentation of various data layers and this was also associated with a new development in the archive:

1. Support for a new Analysis object type 'PROCESSED_READS' has been introduced, which enables the archiving of as yet unsupported data types (here the Tara Oceans Ocean Microbial Reference Gene Catalogue (see below)) derived from the submitted raw reads.
2. The Tara Oceans project PRJEB7988 of the Ocean Microbiome has been registered that represents (a) an Ocean Microbial Reference Gene Catalogue (OMRGC) with more than 40 million non-redundant gene-coding sequences, (b) per-sample gene predictions and (c) per-sample assemblies. The OMRGC and gene predictions are described as Analysis objects with the accession numbers respectively ERZ094224 and ERZ096909-ERZ097151. The assemblies of shotgun metagenomic reads are available as WGS sequence sets and a summary page with an easy navigation to the WGS masters has been created (http://www.ebi.ac.uk/ena/about/tara-oceans-assemblies). The scale of this project required resolving of a number of technical challenges related to scaling up sequence data submission and processing.

The ENA team was also deeply involved in external communication of the Tara Oceans data to the media and public. The release of the five Science articles was preceded by a teleconference with the Science editorial team, external relations and communication officers in EMBL Heidelberg and EMBL-EBI Hinxton and a member of the ENA team. Invited speakers consisted of authors of the research articles and representatives of the databases holding the underlying data, including ENA and PANGAEA. The ENA team was involved in preparation of the flyers, Figure 4, newsletters for EMBL Heidelberg and EMBL-EBI, a science information package for the reporters invited by the Science editor.

The footprint of Tara Oceans data in the ENA was at the time of the Science publications 11.5 terabytes, which is bigger than the text footprint of Wikipedia. This sequence data is associated with a collection of about 7,000 samples, and represents one of the richest, most consistently described contextual data collection in the public domain.



Find more Tara facts and figures at http://bit.ly/1PKByGN

TARA OCEANS
EXPEDITION – THE DATA

361 million km² - the surface area of the world's oceans

140,000 km

The distance sailed by the *Tara* Oceans expedition over 1140 days (2009-2013). Scientists from 40 nations stopped at 210 stations to collect samples.

11,535 gigabytes

The size of the *Tara* datasets in the European Nucleotide Archive as of May 2015. This represents 12,581 gigabases - roughly equivalent to 135 fully sequenced human genomes.

35,000 samples

The scientists on the *Tara* Oceans expedition helped collect, pack and ship around 35,000 samples of plankton and water.

40 million genes

This is the largest DNA sequencing effort ever done for ocean science. The genetic sequences collected could represent tens of thousands of new species (bacteria, single-celled organisms, viruses) and new ecosystem interactions.

Considering the size of the world's oceans, there is much, much more to discover.

7012 datasets

As of May 2015, the European Nucleotide Archive hosts data on 75% of the molecular samples collected – still more are listed in the PANGEA sample registry. This is one of the richest and most consistently described molecular sample collections in the public domain, and it is freely available to everyone.

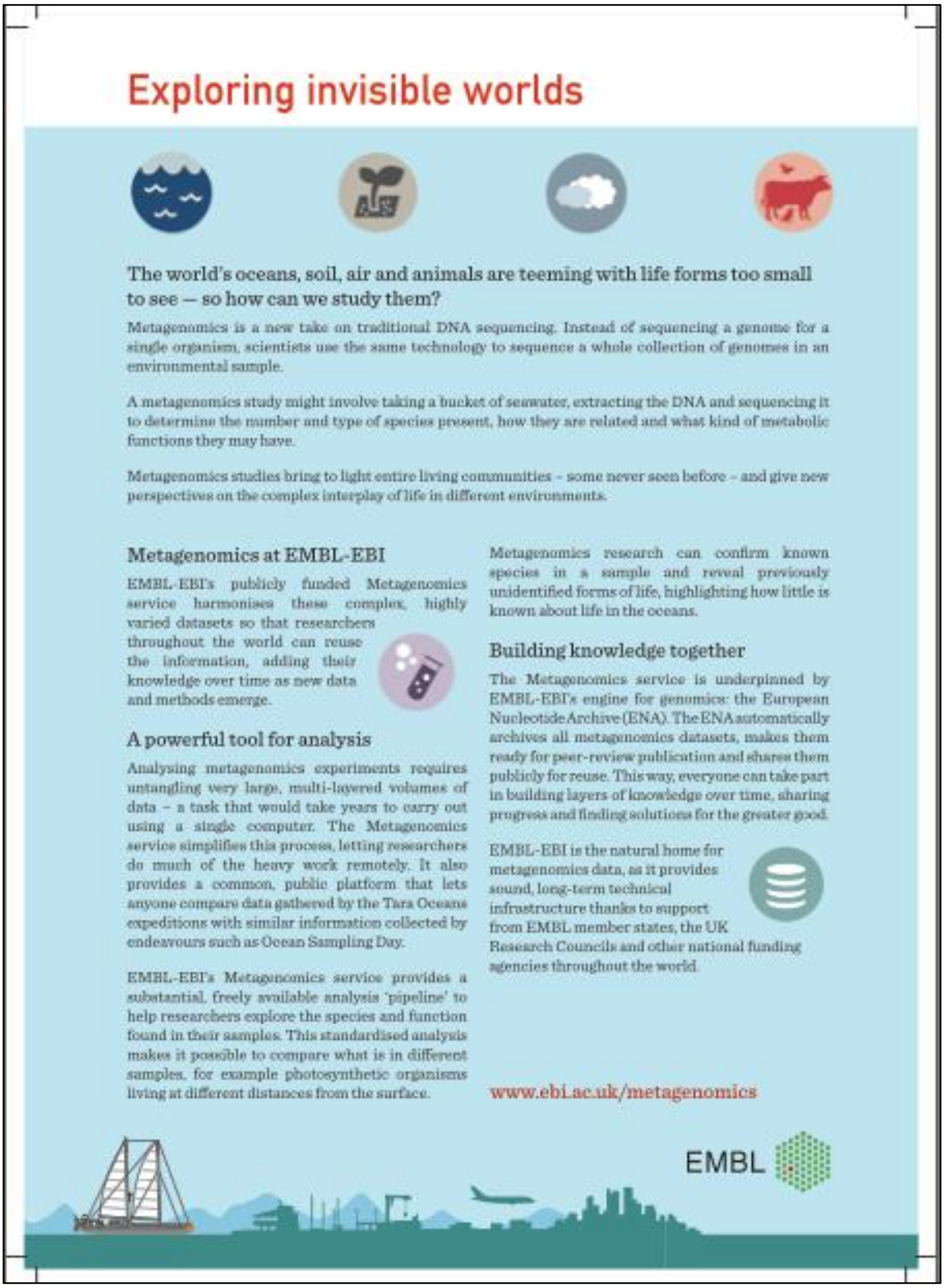
More samples, from *Tara's* Polar Circle expedition, will be submitted in 2015.

Unlimited

Potential to discover new things about life in the world's oceans.

www.ebi.ac.uk/services/tara-oceans-data

(a)



Exploring invisible worlds

The world's oceans, soil, air and animals are teeming with life forms too small to see — so how can we study them?

Metagenomics is a new take on traditional DNA sequencing. Instead of sequencing a genome for a single organism, scientists use the same technology to sequence a whole collection of genomes in an environmental sample.

A metagenomics study might involve taking a bucket of seawater, extracting the DNA and sequencing it to determine the number and type of species present, how they are related and what kind of metabolic functions they may have.

Metagenomics studies bring to light entire living communities – some never seen before – and give new perspectives on the complex interplay of life in different environments.

Metagenomics at EMBL-EBI

EMBL-EBI's publicly funded Metagenomics service harmonises these complex, highly varied datasets so that researchers throughout the world can reuse the information, adding their knowledge over time as new data and methods emerge.

A powerful tool for analysis

Analysing metagenomics experiments requires untangling very large, multi-layered volumes of data – a task that would take years to carry out using a single computer. The Metagenomics service simplifies this process, letting researchers do much of the heavy work remotely. It also provides a common, public platform that lets anyone compare data gathered by the Tara Oceans expeditions with similar information collected by endeavours such as Ocean Sampling Day.

EMBL-EBI's Metagenomics service provides a substantial, freely available analysis 'pipeline' to help researchers explore the species and function found in their samples. This standardised analysis makes it possible to compare what is in different samples, for example photosynthetic organisms living at different distances from the surface.

Metagenomics research can confirm known species in a sample and reveal previously unidentified forms of life, highlighting how little is known about life in the oceans.

Building knowledge together

The Metagenomics service is underpinned by EMBL-EBI's engine for genomics: the European Nucleotide Archive (ENA). The ENA automatically archives all metagenomics datasets, makes them ready for peer-review publication and shares them publicly for reuse. This way, everyone can take part in building layers of knowledge over time, sharing progress and finding solutions for the greater good.

EMBL-EBI is the natural home for metagenomics data, as it provides sound, long-term technical infrastructure thanks to support from EMBL member states, the UK Research Councils and other national funding agencies throughout the world.

www.ebi.ac.uk/metagenomics

(b)

Figure 4a and b: A poster presenting the Tara Oceans project achievements as a part of the press release package preceding five Tara Oceans scientific articles published in Science in May 2015.

All molecular, environmental and bio-imaging data generated by the Tara Oceans project will be accessible from a single Tara Oceans Data Portal, which is currently being developed by the Oceanomics project. EMBL-EBI is contributing to the general design of the portal as well as to correct implementation of links to, and description and retrieval of, individual molecular data sets archived at ENA.

The research vessel Tara moored in September 2015 at Thames Quay in West India Dock South Quay, London en route to the 2015 Paris Climate Congress, COP21. External relations and communication officers in EMBL Heidelberg and EMBL-EBI Hinxton as well as ENA team members were involved once again in preparation and logistics of this special meeting point of scientists and policy makers.

6. Support for Ocean Sampling Day 2014 data presentation

The successful sampling for marine microorganism of the Ocean Sampling Day (OSD) in June 2014 produced a unique collection of 150 datasets generated by in kind contribution of global network of marine stations established during the Micro B3 project. The production of the OSD datasets has been harmonised at all stages: (1) the samples were collected using standardised protocols, (2) metadata have been reported using standardised interface, (3) all DNA extractions have been performed by a single scientist from a single laboratory, (4) all DNA extracts have been sequenced by a single sequencing facility, (5) all analysis has been coordinated by one OSD analysis task group.

In order to provide access to such a unique sampling and sequencing effort the sequencing project PRJEB8682 (http://www.ebi.ac.uk/ena/data/view/PRJEB8682) has been registered at the ENA. This Ocean Sampling Day 2014 authority raw amplicon and metagenome sequencing study represents 150 samples, Figure 3, described according to the M2B3 metadata standard (ten Hoopen *et al.* 2015), associated with 16S and 18S rRNA amplicon sequences and shotgun metagenomes.

Deposition of the OSD datasets to the public molecular data archive enables other data resources to access and add value to the original data. The EBI metagenomics team (EMG) pre-processed the authority raw reads of the project PRJEB8682 and made them available to public in the registered project PRJEB9694 (http://www.ebi.ac.uk/ena/data/view/PRJEB9694). Results of the metagenomic analysis are available from the EBI website at https://www.ebi.ac.uk/metagenomics/projects/ERP009703.

All sequencing and analysis efforts of all OSD data are connected via the registered umbrella project PRJEB5129 (http://www.ebi.ac.uk/ena/data/view/PRJEB5129).

Open access to the OSD data also provided opportunity for the Micro B3 training in the form of the OSD data analysis workshop that took place in March 2015 at the EMBL-EBI in Hinxton, Cambridge, UK. A team of trainers from MPI Bremen, Germany and EMG and ENA from EMBL-EBI, UK, assisted 30 selected researches from the OSD network to learn about

7. Support for sample representation in the GBIF event reports

Microbial diversity discovered through sequencing methods is currently not fully integrated with all biodiversity information available via portals such as the Global Biodiversity Information Facility (GBIF, http://www.gbif.org/). In order to rectify this gap, EMBL-EBI started to explore how ENA environmental sample records be transformed into the Darwin Core format and consumed by GBIF. The ENA and GBIF teams are currently defining the best representation of biodiversity information associated with the metagenomic sample records.

Reference list

Brum, Ignacio-Espinosa, Roux et al. (2015) Patterns and ecological drivers of ocean viral communities. Science 348 (6237), DOI: 10.1126/science.1261498

De Vargas, Audic, Henry, et al. (2015) Eukaryotic plankton diversity in the sunlit ocean. Science 348 (6237), DOI: 10.1126/science.1261605

Lima-Mendez, Faust, Henry et al. (2015) Determinants of community structure in the global plankton interactome. Science 348 (6237), DOI: 10.1126/science.1262073

Sunagawa, Coelho, Chaffron, et al. (2015) Structure and function of the global ocean microbiome. Science 348 (6237), DOI: 10.1126/science.1261359

ten Hoopen, Pesant, Kottman et al. (2015) Marine microbial biodiversity, bioinformatics and biotechnology (M2B3) data reporting and service standards. Standards in Genomic Sciences 10:20, DOI:10.1186/s40793-015-0001-5

Villar, Farrant, Follows et al. (2015) Environmental characteristics of Agulhas rings affect inter-ocean plankton transport. Science 348 (6237), DOI: 10.1126/science.1261447