embric

EUROPEAN MARINE BIOLOGICAL
RESEARCH INFRASTRUCTURE CLUSTER

Deliverable 4.1
Development of data standards for the marine domain

Date: November 2016

HORIZON 2020 - INFRADEV
Implementation and operation of cross-cutting services and solutions
for clusters of ESFRI

Grant agreement no.:	654008
Project acronym:	EMBRIC
Project website:	www.embric.eu
Project full title:	European Marine Biological Research Infrastructure cluster to promote the Bioeconomy
Project start date:	June 2015 (48 months)
Submission due date:	November 2016
Actual submission date:	November 2016
Work Package:	WP 4 – Data services and reporting standards
Lead Beneficiary:	EMBL - EBI
Version:	2.0
Authors:	Petra ten Hoopen, Guy Cochrane, David Smith (MIRRI), Luke Holman, Ian Johnston, Mariella Ferrante

Project funded by the European Union's Horizon 2020 research and innovation programme (2015-2019)

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

Abstract

The EMBRIC deliverable 4.1 aims to develop recommendations on reporting mariculture-related contextual fields making it easier for sampling groups to record and report their data to public data archives and thus to share the data beyond the EMBRIC community. Bioinformatics requirements of use case workflows described in the EMBRIC work packages 6, 7 and 8 were mapped in joint cross-package workshop and followed by discussions with domain experts of each case study. This led to formulation of recommendations for reporting contextual data of marine organisms available in culture collections, specifically microbial and microalgae strains, and recommendations for reporting contextual data of shellfish. These recommendations can be used for reporting contextual data of molecular samples to ELIXIR molecular data repositories.

Table of Contents

1 Introduction

The marine environment with its largely unexplored biodiversity and biotechnological potential provides a challenging opportunity for discovery of organisms with novel properties, attractive for natural product discovery or for selection of exceptional farming strains. The EMBRIC project is an excellent platform for collaboration of specialists and connection of expertise across the EMBRIC cluster of research infrastructures unified in the goal to enhance marine biotechnology.

The main objective of the EMBRIC work package 4 is to provide sustainable data management services for the marine science community. This objective is divided into three tasks, namely:

(1) enable a consultancy service providing bioinformatics configurations for the EMBRIC case studies in the first instance and, in the second phase, for other EMBRIC workflows

(2) establish a data warehouse facilitating better utilisation of bioinformatics resources across the whole EMBRIC cluster

(3) extend existing marine contextual data standard M2B3 for aquacultures enabling simplified data reporting into public repositories and better data integration across public data repositories.

This deliverable 4.1 relates to the development of contextual data standards for the marine domain. The first chapter summarises the EMBRIC case studies workshop aimed at mapping bioinformatics requirements of use cases workflows described in the EMBRIC work packages 6, 7 and 8. The second chapter outlines recommendations for reporting contextual data of marine organisms available in culture collections, specifically microbial and microalgae strains. The third chapter outlines recommendations for reporting contextual data on shellfish.

These recommendations were developed in collaboration with domain experts associated with the workflows of the packages 6,7 and 8. Although considerable effort has been put into the contextual data standardisation, this should be seen as an on going process that requires further refinement of the guidelines as resources and components of the marine bioinformatics infrastructure develop.

2 Mapping EMBRIC case studies requirements workshop

2.1 Workshop objectives

On 2nd and 3rd March 2016 EMBL-EBI organised the EMBRIC case studies requirements workshop. Participants from seven institutes (UiT-Norway, CABI-UK, SZN-Italy, USTAN-UK, FMP-Germany, SB-ROSCOFF-France and EMBL-EBI-UK) discussed bioinformatics needs for workflows described in the EMBRIC Work Packages 6, 7 and 8. In order to support data management of the EMBRIC case studies and future use cases later on, both the current and anticipated **requirements specific to each EMBRIC case study** were discussed covering:

1. **contextual data** reporting including relevant ontologies integration
2. **computational needs** for support of the case study workflows

2.2 Workshop agenda and participants

2.2.1 Agenda

2nd March 2016

Day I – EMBRIC case study-specific needs for data and compute

12.00-13.00 Lunch

13.00-13.15 Welcome and introduction (Guy Cochrane - EMBL-EBI, UK)

13.15-13.45 **Data management support** – extension of the M2B3 data reporting standard and the configurator **(Guy Cochrane/Petra ten Hoopen – EMBL-EBI, UK)**

13.45-14.15 Data management-related discussion and conclusions **(All)**

14.15-14.45 **Microbial workflows** – data and computational needs specific for the use case **(David Smith – CABI, UK)**

14.45-15.15 Microbial workflow needs-related discussion and conclusions **(All)**

15.15-15.30 Coffee break

15.30-16.00 **Microalgae workflows** – data and computational needs specific for the use case **(Mariella Ferrante – SZN, IT)**

16.00-16.30 Microalgae workflow needs-related discussion and conclusions **(All)**

16.30-17.00 **Shellfish and finfish workflows** – data and computational needs specific for the use case **(Ian Johnston – USTAN, UK)**

17.00-17.30 Shellfish/finfish workflow needs-related discussion and conclusions **(All)**

17.30-18.00 Comments and discussion **(All)**

19.00 Dinner – Hinxton Red Lion

3rd March 2016

Day II – meeting case studies needs for data and compute specified on Day I

09.00 Welcome

09.05-09.30 **EU-OPENSCREEN chemical resources** – (overview of existing services and their possible alterations to meet the case studies needs specified on Day I (**Torsten Meiners – FMP, GER**)

09.30-09.45 Discussion

09.45-10.15 **ELIXIR chemical resources** – (overview of existing services and their possible alterations to meet the case studies needs specified on Day I (**Reza Salek – EMBL-EBI, UK**)

10.15-10.30 Discussion

10.30-10.45 *Coffee break*

10.45-11.15 **ELIXIR genomic resources and cloud compute** – (overview of existing services and their possible alterations to meet the case studies needs specified on Day I (**Guy Cochrane – EMBL-EBI, UK**)

11.15-11.30 Discussion

11.30-12.45 Summary of conclusions and planning next steps **(Guy Cochrane/Petra ten Hoopen – EMBL-EBI, UK)**

12.45-13.00 Closure of meeting

13.00 *Lunch*

2.2.2 Participants

The following EMBRIC partner institutes participated in the Use Case Workshop:

- Ian Johnston (USTAN, UK, <iaj@st-andrews.ac.uk>)
- Alicia Bertolotti (USTAN, UK, <r01acb15@abdn.ac.uk>)
- David Smith (CABI, UK, <d.smith@cabi.org>)
- Rebecca Gross (USTAN, UK, <rjmg@st-andrews.ac.uk>)
- Mariella Ferrante (SZN, IT, <mariella.ferrante@szn.it>)
- Torsten Meiners (FMP, GER, <meiners@fmp-berlin.de>)
- Martin Neuenschwander (FMP, GER, <neuenschwander@fmp-berlin.de>)
- Mark Hoebeke (SB-ROSCOFF, FR, <mark.hoebeke@sb-roscoff.fr>)
- Reza Salek (EMBL-EBI, UK, <reza.salek@ebi.ac.uk>)
- Nils Peder Willassen (UiT, NO, <nils-peder.willassen@uit.no>)
- Guy Cochrane EBI (EMBL-EBI, UK, <cochrane@ebi.ac.uk>)
- Petra ten Hoopen (EMBL-EBI, UK, <petra@ebi.ac.uk>)

2.3 Data management support

2.3.1 EMBRIC WP4: Data services and reporting standards by Guy Cochrane

Guy focused in his talk on the EMBRIC Configurator. He explained that the configurator service should be seen as a service that takes a project proposal and turns it into a business plan for the project. WP4 will support EMBRIC use cases to create their case-specific configurations but will also run the configuration (i.e. help to execute the business plan). This will create a knowledgebase for development of configurations for other marine campaigns in the future.

Discussion related to the fact that the microbial strains use case represents only a subset of technologies available and this has to be taken into account in the data management support.

2.3.2 EMBRIC Data reporting standards by Petra ten Hoopen

Petra explained in her talk the M2B3 Data Reporting Standard developed in the frame of the Micro B3 project and provided two examples of marine campaigns, Tara Oceans and Ocean Sampling Day, where contextual data in the primary data archives (PANGAEA and European Nucleotide Archive) are described using the M2B3 Standard. She focused on the legacy of the M2B3 and its potential for the EMBRIC community. Petra then drafted an extension of the M2B3 for microbial strains, microalgae, shellfish, finfish and bioactive compounds in order to invite meeting participants to comment.

Discussion related to:

- Suitability of the MIRRI Minimal Data Set for a strain identification but not for a promising strains selection.
- Revision of descriptors captured by microalgae collections to get a better idea of their relevance.
- Feasibility to select traits mostly relevant for shellfish/finfish breeders.
- Distinction of attributes by measurement from attributes by inference

2.4 EMBRIC microbial workflows

2.4.1 Microbial workflow – data and computational needs specific for the use case (intrinsically linked to RI information systems) by David Smith

David explained the aim of going towards maximum value from marine microbial resources. He presented bottlenecks, actions to address them and problems related to this. David then focused on parallel pipelines for genomics and metabolomics, from a strain selection, via generation of data from primary screening (identifying candidate metabolites and clusters) to further material selection (using heterologous expression) and back to data on structural characterization and compounds. In the second part David presented outputs reached so far by MIRRI and WP6 requirements to support planned steps of the workflow, where the need for machine-readable access to published methods appeared recurrently.

Discussion related to:

- challenge to select organisms of potential from the vast amount of as yet uncultured species
- strain selection method predetermines which strains will be selected; for instance, a biofilm or iChip will preselect only certain strains that form biofilms or that match the iChip requirements
- WP3 and WP6 review of available technologies, which will help to decide which technology to use to obtain strains of interest
- challenge of identifying the potential of strains available in culture collections and the need for data integration where chemical pathways can be linked to strain identifications to better reveal their potential; Normal practice of culture collections is to publish generated sequences but this practice is not established for metabolic profiles

2.5 EMBRIC microalgae workflows

2.5.1 Microbial workflow for blue biotechnological application by Mariella Ferrante

Mariella clarified the high potential of microalgae as a source of natural products and as a target for breeding and presented the package workflow including species selection, metabolic profiling, cell screening of promising fractions and bioactive compounds characterisation. Afterwards, Mariella specified partners responsible for each task and relevant arrays, and highlighted that the anti-inflammatory screening will be given priority due to high costs of other bioactive screenings (such as anticancer, antioxidant). The main need of the WP7 is a resource enabling deposition, search and retrieval of data on microalgae strains, their mutants and their compounds.

Discussion related to:

- need for a resource that would archive in a structured way protocols describing biological resource handling and analysis. Would www.protocol.io be an option?
- active support of ABS by data resources by having the established practice of recording the sampled material provenance

2.6 EMBRIC shellfish and finfish workflows

2.6.1 Genetics and selective breeding in aquaculture species by Ian Johnston

Ian explained the fundamental equation of breeding, where phenotype is a result of genotype responding to its surrounding environment, outlined the traditional breeding scheme, listed the most important traits (including growth rate, mortality, quality and fecundity) and stressed that the traits collection should be automatic and done in an affordable way. Despite existence of relevant ontologies, ATOL and EOL, each breeding company uses a specific trait terminology. However, a new culture of increasing collaboration is emerging. Ian then focused on use of genetic resources for breeding of aquaculture species, provided an example of a breeding experiment, targeted species and compute needs.

Alicia explained her workflow for study of copy number variation in salmon.

2.7 EU-OPENSCREEN chemical resources

2.7.1 EU-OPENSCREEN – chemical tools for the life sciences by Torsten Meiners

Torsten introduced the transition phase of EU-OPENSCREEN, its position on the scale of the intellectual property value as well as the main outputs of the RI being chemical biology data and compounds for bioactive entity screening, thus not only a repository for well-characterised drug compounds. ECBD (European Chemical Biology Database) will be a portal to this unique compound collection, using CHEBI identifiers and links to major repositories for targets, such as ATCC collection, NCBI Taxonomy, UniProt, BRENDA ontology. Compounds and Assays (specified by developed BioAssay ontology) are provided by users. Expertise for natural product discovery is fragmented across RIs and there is a high need for access to harmonised multidisciplinary workflow.

Discussion related to:

- Establishing ECBD links from compound records to BioSample identifiers which would link the compound with its environmental provenance
- CHEMBL chemical data deposition guide that better clarifies ECBD reporting requirements and will be relevant for the EMBRIC M2B3 extension.

2.8 ELIXIR chemical resources

2.8.1 MetaboLights: Capture and dissemination of metabolomics data by Reza Salek

Reza did not feel well and apologised for not attending the second day of the meeting in person. However, Reza presented his talk via a Skype call. His dedication was much appreciated by all workshop participants. Reza focused on introducing the MetaboLights resources as a relevant resource for depositing metabolite profiles produced in the EMBRIC microbial and microalgae workflows. Reza explained the ISA data model and submission tool and showed examples of existing MetaboLight records.

Discussion related to:

- Comments on usefulness of this resource for their respective work package.
- possibility of cross-referencing identified metabolites to compounds in the ECBD database

2.9 ELIXIR genomic resources and cloud compute

2.9.1 ELIXIR genomics and cloud resources by Guy Cochrane

Guy very briefly introduced the ELIXIR RI and then provided a quick tour through the major ELIXIR genomics resources hosted by the EMBL-EBI: European Nucleotide Archive, Ensembl, UniProt, Expression Atlas, EBI Metagenomics Portal, MetaboLights, EMBL-EBI BioSample Database. Guy then explained the value of the NCBI Taxonomy as the unified taxonomic index used across most EMBl-EBI resources. Afterward, Guy clarified how cross-references to data and physical material in biorepositories are handled at the ENA. In the last part of his talk Guy provided an insight into the ELIXIR compute services and tools.

Discussion related to:

- possibilities of ELIXIR experts delivering training at the partner site or possibilities for EMBRIC partners to visit the ELIXIR hub.

3 EMBRIC recommendation for reporting contextual data of molecular samples from microbial culture collections

3.1 Contextual data checklist for a molecular sample from culture collection strains

Microbial domain Biological Resource Centre (mBRC) or culture collection communities have practices established over decades- for the description of microorganisms available in their collections. Requirements on data formats and minimal data sets vary between repositories and depend also on the taxonomic group of the deposited culture. The OECD Best Practice for Biological Resource Centres http://www.oecd.org/sti/biotech/oecdbestpracticeguidelinesforbiologicalresourcecentres.htm attempts to address this and to define data sets for national culture collections, such as the UK Culture Collection Organisation, for projects, such as CABI, and for the **E**uropean Consortium of **M**icro**b**i**a**l **R**esources **C**entres – EMbaRC. The EMbaRC operational standard specifies Minimum Data Set (MDS), Recommended Data Set (RDS) and Full Data Set (FDS) for each of the following taxonomic groups: bacteria, archaea, cyanobacteria, fungi, protozoa, microalgae, yeast, virus and phage http://www.embarc.eu/deliverables/EMbaRC_D.NA1.1.2_D2.37_Data_Std.pdf.

The Genomic Standards Consortium initiative formulated Minimum information about Genome Sequence (MIGS)(Yilmaz, 2011) as part of the unified standard Minimum information about any (x) Sequence (MIxS). Although highly relevant for description of contextual data of genomes, this standard is more suitable for environmental samples, where minimal required information include georeference, collection date, environment biome, feature or material. However, this is frequently not reported in cultured collections and molecular samples from such culture collection cannot be searched based on of these descriptors.

Similarly, the M2B2 standard of minimal information about marine microbial sample (ten Hoopen, 2015) is tailored to marine environmental samples and enables interdisciplinary interoperability of genomic, oceanographic and biodiversity data generated from these marine microbial samples.

The main objective of the EMBRIC recommendation for reporting contextual data of a molecular sample from a culture collection strain is to identify descriptors of the microbial strains deposited in mBRCs, which shall be consistently reported in molecular data archives in order to be useful to scientists for discovery of associated molecular data.

Inevitably, the EMBRIC recommendations draw from the EMbaRC recommendations but by no means wish to replicate all information available in mBRCs.

The contextual data checklist for molecular samples originating from culture collection strains, **Table 1 A-C**, represents core descriptors of the samples that will allow the strain to be identified in mBRCs and additional information, retained in the mBRCs, to be found. This should also enable useful search of molecular data associated with the sampled strain. The concept of strain can have different meaning in different collections. It is expected that each collection hold definition of the concept for its strains. The granularity of strain descriptors in molecular repositories shall form a balance between too many, which creates difficulties in adopting the recommendations by the relevant expert community, and too few, which can affect the molecular data discovery. Descriptors of the checklist are divided into three categories: mandatory (required for molecular samples of all microorganisms from collections), recommended (highly relevant to samples of some organism types) and optional (relevant to samples of some organism types).

Table 1 A-C summarises for each descriptor its name, definition, requirement level, format and example. Microalgae is used as an example here. Other categories of microorganisms in mBRCs include bacteria, archaea, cyanobacteria, fungi, protozoa, yeast, virus and phage.

Asterisk at the descriptor name indicates its current availability for genomic data search in the public genomic data archive, the European Nucleotide Archive. This indexing example demonstrates how specific genomic data subsets can be discovered using combinations of the indexed descriptors.

Table 1 Information about a biological sample that originates from a microbial strain deposited in a microbial domain Biological Resource Centre (mBRC) and is associated with molecular data (e.g. genomic or metabolomic). **A** – minimal information, mandatory for any molecular sample from a cultured collection; **B** – recommended information, applicable and highly relevant for some organism types; **C** – optional information, applicable and relevant for some organism types.

1A

descriptor name	**descriptor definition**	**descriptor Requirement level**	**descriptor format**	**example**
sample ID *	unique identifier for the sample	mandatory	Single-line text	lab barcode XY
sample title *	a brief human readable description of the sample	mandatory	Single-line text	Sample obtained from the 9A progeny strain of parent strains 88 and 75. This sample has a biological replica XZ.
organism scientific name *	scientific name of the organism in the culture	mandatory	NCBI Taxonomy ID	Seminavis robusta

culture collection *	institution code and identifier for the culture from which the sample was obtained, with optional collection code.	mandatory	Single-line text	DCG 0096
organism type	type of the organism in the culture	mandatory	Single-line text controlled by a list of allowed values: bacteria, archaea, cyanobacteria, fungi, protozoa, microalgae, yeast, virus, phage.	microalgae
growth condition	A role that a material entity can play which enables particular conditions used to grow organisms or parts of the organism. This includes isolated environments such as cultures and open environments such as field studies	mandatory	Single-line text	medium: f/2+Si

1B

descriptor name	**descriptor definition**	**descriptor requirement level**	**descriptor format**	**example**
WDCM registry number	unique number of the mBRC approved by the World Data Centre for Microorganisms (WDCM)	recommended	Single-line text	WDCM 1039
organism name synonym	synonym or other name of the organism in the culture; an alternative taxonomy can be used here, such as DSMZ or MycoBank Taxonomy	recommended	Single-line text	Not provided
strain *	name of strain from which sample was obtained	recommended	Single-line text	9A
serotype *	serological variety of a species characterized by its antigenic properties	recommended	Single-line text	Not applicable
serovar *	serological variety of a species (usually a prokaryote) characterized by its antigenic properties	recommended	Single-line text	Not applicable
pathotype	name or code for pathotype of organism	recommended	Single-line text	Not applicable
host taxid	NCBI taxon id of the host, e.g. 9606	recommended	Single-line text	Not applicable
geographic location (country and/or sea) *	the geographical origin of the sample as defined by the country or sea; country or sea names should be chosen from the INSDC country list (http://insdc.org/country.html)	recommended	Single-line text controlled by a list of allowed values: INSDC country list	The Netherlands
geographic location	the geographical origin of the sample as	recommended	Single-line	Zeeland, "Veerse

(region and locality)	defined by the specific region name followed by the locality name		text	Meer" lake
geographic location (latitude) *	the geographical origin of the sample as defined by latitude and longitude; the values should be reported in decimal degrees and in WGS84 system	recommended	DD, WGS 84 for GPS	51.543333
geographic location (longitude) *	the geographical origin of the sample as defined by latitude and longitude; the values should be reported in decimal degrees and in WGS84 system	recommended	DD, WGS 84 for GPS	3.804167

1C

descriptor name	**descriptor definition**	**descriptor requirement level**	**descriptor format**	**example**
collection date *	the date of sampling, either as an instance (single point in time) or interval. In case no exact time is available, the date/time can be right truncated i.e. all of these are valid ISO8601 compliant times: 2008-01-23T19:23:10+00:00; 2008-01-23T19:23:10; 2008-01-23; 2008-01; 2008.	optional	Single-line text	2015-06-15
collected by *	name of persons or institute who collected the specimen	optional	Single-line text	T. Moons
genotype	name or code for genotype of organism	optional	Single-line text	Not provided
organism phenotype	where possible, please use the Experimental Factor Ontology (EFO) to describe your phenotypes.	optional	Single-line text	average diameter-65 µm
propagation	this field is specific to different taxa. For phages: lytic/lysogenic, for plasmids: incompatibility group (Note: there is the strong opinion to name phage propagation obligately lytic or temperate, therefore we also give this choice. Mandatory for MIGS of eukaryotes, plasmids and viruses.	optional	Single-line text	isogamy auxosporulation??
developmental stage *	if the sample was obtained from an organism in a specific developmental stage, it is specified with this qualifier	optional	Single-line text	Not provided
sample storage duration	duration for which sample was stored	optional	Single-line text	Not provided
sample storage temperature	temperature at which sample was stored, e.g. -80	optional	Single-line text	Not provided
sample storage location	location at which sample was stored, usually name of a specific freezer/room	optional	Single-line text	Not provided

4 EMBRIC recommendation for reporting contextual data of molecular samples from shellfish

4.1 Contextual data checklist for a molecular sample from shellfish

Aquaculture is a fast growing sector of agriculture. Shellfish contribute significantly to the overall production although only a few species from the hundreds of thousands known shellfish species have been cultured so far. Next to conventional breeding programmes a marker-assisted selection is a very promising area of aquaculture research enabling brooders to be selected according to both genotypes and performance. Marker-assisted selection requires DNA markers that should ideally be the causative mutation underlying the phenotypic variation.

Consistent and accurate recording of intrinsic and environmental traits associated with shellfish genomic data is a prerequisite to linking its genotype and phenotype.

The main objective of the EMBRIC recommendation for reporting contextual data of a shellfish molecular sample is to encourage consistent reporting of minimal shellfish contextual information in molecular data archives that would be useful to scientists for discovery of associated molecular data.

The contextual data checklist for a molecular sample from shellfish has been developed in collaboration with shellfish aquaculture experts associated with the EMBRIC work package 8. Descriptors of the checklist are divided into three categories: mandatory (required for molecular samples from all shellfish), recommended (highly relevant to some shellfish samples) and optional (relevant to some shellfish samples). Table 2 summarises for each descriptor its name, definition, requirement level, format and example.

Table 2 Information about a shellfish biological sample that is associated with molecular data. **A** – minimal information, mandatory for all shellfish molecular samples; **B** – recommended information, applicable and highly relevant to some shellfish samples; C – optional information, applicable and relevant for some shellfish samples.

Asterisk at the descriptor name indicates its current availability for genomic data search in the public genomic data archive, the European Nucleotide Archive. This indexing example demonstrates how specific genomic data subsets can be discovered using combinations of the indexed descriptors.

2A

descriptor name	descriptor definition	descriptor Requirement level	descriptor format	example
sample ID*	unique identifier for the sample	mandatory	Single-line text	lab barcode XY
sample title*	a brief human readable description of the sample	mandatory	Single-line text	Sample obtained from the 9A progeny strain of parent strains 88 and 75. This sample has a biological replica XZ.
organism scientific name*	scientific name of the organism in the culture	mandatory	NCBI Taxonomy ID	*Pecten maximus* (taxid:6579)
sampling campaign*	refers to a finite or indefinite activity aiming at collecting data/samples, e.g. a cruise, a time series, a mesocosm experiment.	mandatory	Single-line text	TARA_20110401Z.
sampling station*	refers to the site/station where data/sample collection is performed.	mandatory	Single-line text	TARA_100.
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_bodc_vocab_v2/search.asp?lib=L06, and Platform name.	mandatory	Single-line text	Research Vessel Tara
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 event. Format: yyyy-mm- ddThh:mm:ssZ	mandatory	Single-line text	2013-06-21T14:05:00Z/2013-06-21T14:46:00Z
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	mandatory	Single-line text	-24.6666
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	mandatory	Single-line text	-096.1012
depth*	the distance below the surface of the water at which a measurement was made or a sample was collected. Format: ####.##, Positive below the sea surface. SDN:P06:46:ULAA for m.	mandatory	Single-line text	14.71

protocol label*	identifies the protocol used to produce the sample, e.g. filtration and preservation	mandatory	Single-line text	BACT_NUC_W0.22-1.6
environment biome*	biomes are defined based on factors such as plant structures, leaf types, plant spacing, and other factors like climate. Biome should be treated as the descriptor of the broad ecological context of a sample. Examples include: desert, taiga, deciduous woodland, or coral reef. EnvO (v 2013-06-14) terms can be found via the link: www.environmentontology.org/Browse-EnvO	mandatory	Single-line text	marine biome (ENVO: 00000447)
environment feature*	environmental feature level includes geographic environmental features. Compared to biome, feature is a descriptor of the more local environment. Examples include: harbor, cliff, or lake. EnvO (v 2013-06-14) terms can be found via the link: www.environmentontology.org/Browse-EnvO	mandatory	Single-line text	sea grass bed (ENVO: 01000059)
environment material*	the environmental material level refers to the material that was displaced by the sample, or material in which a sample was embedded, prior to the sampling event. Environmental material terms are generally mass nouns. Examples include: air, soil, or water. EnvO (v 2013-06-14) terms can be found via the link: www.environmentontology.org/Browse-EnvO	mandatory	Single-line text	cobble sediment (ENVO: 01000115)
seabed habitat	classification of the seabed where the organism has been found; for European seabed habitats please use terms from http://eunis.eea.europa.eu/habitats-code-browser.jsp;	mandatory	Single-line text	B3.4 : Soft sea-cliffs, often vegetated
age	age of the organism the sample was derived from	mandatory	Single-line text	2 months
aquaculture origin	origin of stock and raised conditions	mandatory	Single-line text controlled by a list of allowed values: AOAR, WOAR, WOWR	WOAR (Wild Origin
shellfish total weight	total weight of shellfish including shell at the time of sampling. Epifauna and epiphytes to be removed	mandatory	Single-line text	223g
shellfish soft tissue	total weight of all soft tissue, i.e. weight	mandatory	Single-line text	83g

weight	of entire organism without shell, at the time of sampling			
shell length	length of shell (perpendicular to the hinge)	mandatory	Single-line text	123mm
shell width	width of shell (perpendicular angle to length)	mandatory	Single-line text	110mm

2B

descriptor name	descriptor definition	descriptor Requirement level	descriptor format	example
adductor weight	total weight of striated muscle and smooth muscle	recommended	Single-line text	33.2g
gonad weight	total weight of entire gonad tissue	recommended	Single-line text	6.7g
shell markings	visible markings on outer shell	recommended	Single-line text	dark striations
toxin burden	concentration of toxins in the organism at the time of sampling	recommended	Single-line text	502mg/kg
marine region	the geographical origin of the sample as defined by the marine region name chosen from the Marine Regions vocabulary at http://www.marineregions.org/.	recommended	Single-line text	Adriatic Sea (MRGID:3314)

2C

descriptor name	descriptor definition	descriptor Requirement level	descriptor format	example
sample collection device	the sampling device(s) used for the Event.	optional	Single-line text	CTD(sbe9C)/Rosette with Niskin bottles
storage conditions (fresh/frozen/other)	explain how and for how long the soil sample was stored before DNA extraction	optional	Single-line text	-80 degree Celsius, 1month
sample health state	health status of the subject at the time of sample collection	optional	Single-line text controlled by a list of allowed values: healthy, diseased	diseased
sample disease status	list of diseases with which the subject has been diagnosed at the time of sample collection; can include multiple diagnoses;	optional	Single-line text	Vibrio spp.

	the value of the field depends on subject;			
treatment agent	the name of the treatment agent used	optional	Single-line text	antibiotics
chemical compound	a drug, solvent, chemical, etc., with a property that can be measured such as concentration (http://purl.obolibrary.org/obo/CHEBI_37577).	optional	Single-line text	oxytetracycline (CHEBI:27701)

Implementation of the shellfish contextual data checklist described in the Table 2 A-C is available from the European Nucleotide Archive [xx] and will serve for deposition of shellfish molecular data to the ELIXIR data resources.

Unlike standardisation of shellfish contextual data, the selection of relevant traits for finfish is far more advanced and specific ontologies have been developed, such as the Animal Trait Ontology for Livestock (ATOL) and Environment Ontology for Livestock (EOL) developed at INRA well suitable for finfish traits description. However, finfish breeding companies frequently use their specific trait terminology and traits selection. This hinders further development in this area significantly and calls for a new culture of increasing collaboration that can bring new opportunities to further develop and make available to public bioinformatics services for aquaculture of finfish.

5 Conclusion

This EMBRIC deliverable 4.1 main objective is to make available recommendations on reporting mariculture-related contextual fields making it easier for sampling groups to record and report their data to public data archives and thus to share their data beyond the EMBRIC community.

In order to understand specific bioinformatics needs of the scientific community associated with the EMBRIC project, we have organised a joint workshop of the EMBRIC work package delivering the data management services (WP4) with work packages representing the EMBRIC case studies (WP6, WP7 and WP8). Details of this workshop are summarised in the chapter 2 of this report.

Following on from the workshop discussions we have established communication channels with each of the case studies and in collaborations with their experts formulated recommendations for reporting contextual information of molecular samples originating from cultured collections. These are summarised in the chapter 3.

Similarly, we have identified in series of e-meetings with aquaculture domain experts minimal contextual fields that shall accompany molecular data of shellfish. The shellfish contextual data checklist is now also implemented in the submission system Webin of the European Nucleotide Archive, the European partner of the global INSDC archive of public nucleotide sequence data. This tool is publicly available to any research group depositing sequence data to the ELIXIR data resources.

6 References

[1] OECD Best Practice Guidelines for Biological Resource Centres; http://www.oecd.org/sti/biotech/oecdbestpracticeguidelinesforbiologicalresourcecentres.htm

[2] Yilmaz P, Kottman R, Field D, Knight R, Cole JR, Amaral-Zettler L et al. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nature Biotechnology 2011;29:415-20.

[3] ten Hoopen P, Pesant S, Kottmann R, Kopf A, Bicak M, Claus S et al. Marine microbial biodiversity, bioinformatics and biotechnology (M2B3) data reporting and service standards. Standards in Genomic Sciences 2015;10:20

[4] ENA shellfish checklist, https://www.ebi.ac.uk/ena/data/view/ERC000036

[5] Animal Trait Ontology for Livestock (ATOL), Environment Trait Ontology for Livestock (EOL), http://www.atol-ontology.com/rb/en/1