Joint European Research Infrastructure network for Coastal Observatory – Novel European eXpertise for coastal observaTories - **JERICO-NEXT**	
Deliverable title	Marine biological data: quality control and management practices
Work Package Title	WP5: Data Management
Task title	Integration of biological data
Deliverable number	D5.4
Description	This document describes the biological DMP including an elaborated section on quality control procedures for biological data.
Lead beneficiary	VLIZ – Flanders Marine Institute
Lead Authors	L. Tyberghein, S. Claus, K. Deneudt
Contributors	
Submitted by	
Revision number	V4.0
Revision Date	20/02/2017
Security	Public

The JERICO-NEXT project is funded by the European Commission's H2020 Framework Programme under grant agreement No. 654410
Project coordinator: Ifremer, France.

History			
Revision	**Date**	**Modification**	**Author**
V1	25/08/2016	First draft	LT
V2	12/09/2016	update	LT, SC
V3	15/09/2016	update	LT, SC, KD
V4	20/02/2017	update	LT, SC

Approvals				
	Name	**Organisation**	**Date**	**Visa**
Coordinator	Patrick Farcy	Ifremer	27/02/2017	PF
WP Leaders	Leonidas Perivoliotis	HCMR	27.02.2017	LP

Table of contents

1. Introduction

In JERICO-NEXT, a major effort is initiated towards a stronger integration of biological information as part of the observing networks. These objectives will be realized through the integration of mature technologies and the development of emerging technologies capable of delivering operational biological data.

During the recent years, large marine biological data systems have been created to store, archive and integrate traditional marine biological data. In the framework of JERICO-NEXT an operational link will be created with EMODnet biology, the biological component of the European Marine Observation and Data Network and OBIS, the Ocean Biogeographic Information System. This will facilitate the data exchange between the coastal observatories and the existing marine biological data networks and hereby maximize access to marine biological data for any type of user on a marine basin- or region-wide basis.

Integration of data involves the implementation of data quality control steps. Evaluating and documenting data quality is since many years standard practice in disciplines like medicine and genetics, however only the last years similar effort is being done for biological data, more particular for biodiversity data.

This report is part of the JERICO-NEXT WP5 on data management. The first objective is the description of the general biological data management practices. This aims at providing procedures and methodologies to enable data collected through the project to comply with the international standards regarding their quality and metadata. The second objective is to focus on the details of biological data quality control.

This report is seen as a living document that can be amended during the course of the project to incorporate progressive insights or in light of specific emerging needs.

2. Main report

2.1. Biological data management practices

The figure below shows the JERICO-NEXT biological data flow. The idea is to create an interoperable data flow that starts close to the in situ measurements (sample and sensor data), includes archival in archiving data centers and facilitates uptake or redistribution through the relevant European and global databases EMODnet and OBIS.

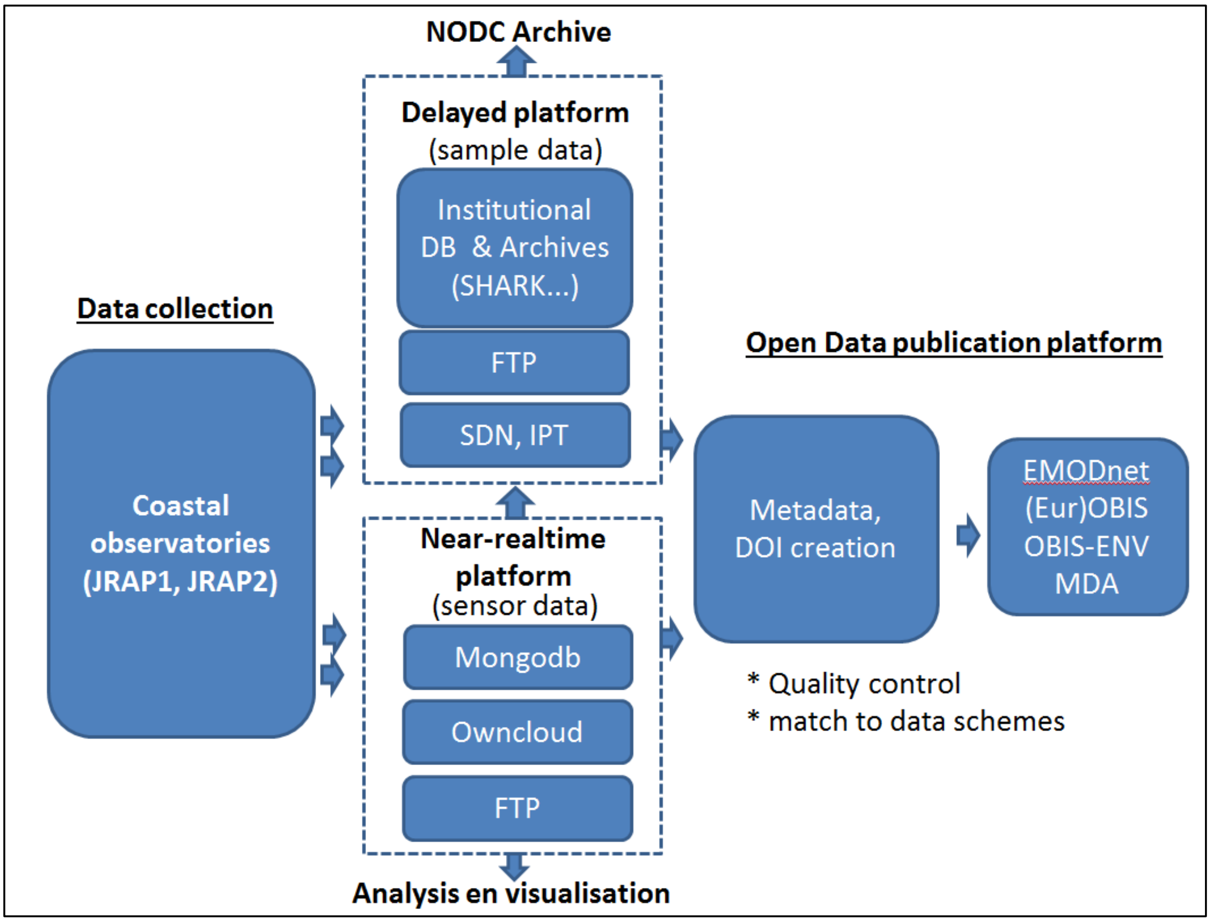

Figure 1: The biological data flow consists of a series of steps describing the path from data collection to data publication.

2.1.1. Data collection

JERICO-NEXT biological data will be collected in the framework of WP3 and WP4 (JRAP1 and JRAP2) activities. At the start of the project a data information survey was performed among the JRAP leaders to provide a clear view of the diversity of the biological data that are expected to be collected during these activities. The document illustrated the diverse types of data that will result from this scientific work. Currently, a large proportion of the JERICO-NEXT biological data fits in the existing configuration of EMODnet Biology. For the additional part, alternatives should be explored.

2.1.2. Data provided to a delayed or Near-Real time storage system

The data can be provided to both a delayed or Near-Real Time (NRT) storage system of choice. The delayed storage systems (mostly containing data from more traditional grab or net sampling and identification activities) are generally the local servers at the collectors site (e.g. the SHARK database from SMHI) and may or may not be connected to a more distributed system like SeaDataNet (SDN). NRT storage systems contain mostly continuous sensor data and these data can be provided on ftp servers, through dedicated databases or web services. As part of WP5 activities it is the intention to connect these data sources by picking up the data flow from either one of these storage systems. Where no systems are in place yet, possibilities for integration and archiving data will be offered. The biological data management approach will be in agreement with international data standards and conventions (see below).

2.1.3. Metadata description

In order to create a digital inventory of the data resources created in the framework of JERICO-NEXT, each of the partners will describe their datasets in a digital catalogue, using an online form. The scope, characteristics, state and accessibility of the data will be documented following common standardized formats based on **ISO19115 metadata standard**. In case the data is already accessible through local online databases a web link to the existing interfaces will be included in the dataset description. The underlying system, linked to the EMODnet biology dataset catalogue, can be plugged into the JERICO-NEXT website and can allow users to query and select data resources based on a set of relevant search criteria. To make the collected data resources traceable and citable, JERICO-Next partners have the opportunity to formally publish their datasets by the assignment of Digital Object Identifiers (DOIs). The implementation of DOIs to track JERICO-NEXT resources used in scientific publications or reports can then be used to demonstrate the impact of the project. JRAP leaders should encourage the upload of metadata in the system as described above. It will require the necessary interaction between the dataset collectors and the data catalogue managers to make this a smooth process. VLIZ will actively ask for metadata information and provide support for this activities.

Links for metadata submission:

- http://emodnet-biology.eu/contribute (click advanced)

2.1.4. Dataset screening

After the metadata description datasets are curated and subjected to screening procedure. This exercise allows to see which data are fit to upload in the existing databases. The biological data collection in JERICO-NEXT will contain a diverse set of biological variables. We will make use of common data schemas and standards to integrate data in the existing systems like **EMODnet** and **OBIS**. In addition, biological data derived from sensor readings will be analyzed and the possibility to develop new schemas or alternative data integration methods will be explored.

2.1.4.1. Common data schemas

JERICO-NEXT biodiversity data will be integrated using the standard procedures adopted by (Eur)OBIS and EMODnet. These systems are well interconnected and the data flow is well established (see figure below and explanation: http://www.eurobis.org/data_flow).

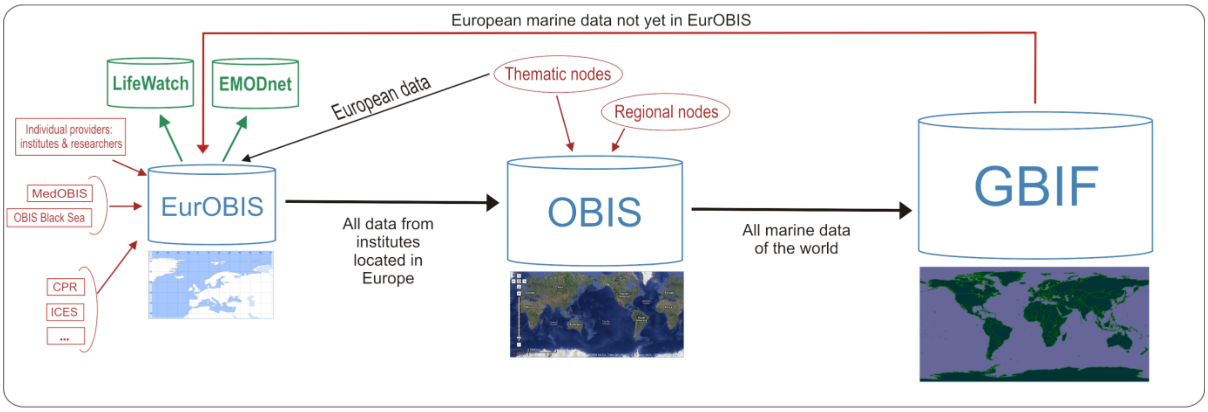

Figure 2: International biodiversity data flow.

Within these international data systems the **Darwin Core standard** is used. This is the standard for biodiversity informatics and it provides stable terms and vocabularies for sharing biodiversity data. Darwin Core is maintained by TDWG (Biodiversity Information Standards).

Description of this standard can be found at:

- http://rs.tdwg.org/dwc/
- http://iobis.org/manual/darwincore/
- http://www.eurobis.org/data_formats

2.1.4.2. Common data formats

Biological data exchange within JERICO-NEXT will happen using the **by EMODnet recommended data formats**. Biological data can be delivered in common data file formats as spreadsheets, relational databases or simple text files.

Next to that biological data can be published through **IPT, Integrated Publishing Toolkit**. IPT is used to publish and share biodiversity datasets and is specifically designed for interoperability: it enables the publishing of content in databases, Microsoft Excel spreadsheets, or text files using open standards namely the Darwin Core and the Ecological Metadata Language.

More information:

- Guidelines for installing: http://iobis.org/manual/ipt/
- Example of ipt server: http://ipt.vliz.be/

There is currently an operational link between EMODnet Biology and the **SeaDataNet** infrastructure. During SDN II action was carried out to make SeaDataNet better fit for handling marine biological data sets and establishing interoperability with biology infrastructure developments. Based on an analysis of present biology data standards and initiatives, such as the OBIS, GBIF, TDWG and WoRMS standards, a recommended format for data transport of biological data was developed. This format deviates from the classical ODV format as the data is stored semantically as is the case in Darwin Core. The format enables partners to make their biological data accessible using the SDN infrastructure and this also enables use SeaDataNet to exchange biological data and to contribute to EMODnet.

More information (incl. guidelines and template) following this links:

- http://www.seadatanet.org/Standards-Software/Data-Transport-Formats
- http://www.seadatanet.org/content/download/27948/190493/file/SDN2_D84b_WP8_ODV_biology_variant_format%20guidelines.pdf

2.1.4.3. Quality control procedures

Section 2.2 addresses the topic of QC procedures for biological data management in detail.

2.1.4.4. Exploration of new data schemas

Biological data will be provided by WP3 and WP4. The biological part of these work packages focuses on the use of innovative observation techniques for addressing pelagic and benthic diversity questions. At present, these innovative techniques and resulting data have not received widespread usage among EU projects for their management and quality control according to international standards. In this respect new possibilities for integration of data will have to be explored in close collaboration with WP3 and WP4. This is currently ongoing in the framework of dedicated organized workshops and meetings. Annex 1 gives an overview of the biological parameters that can be derived from the data sources (sensors and other sampling) in JRAP1 and JRAP2. It is also indicated what parameters can already be included using the current standards and which not.

Several ongoing and booting projects also envision the incorporation of similar data sources. It is the intention to line up with these. The SeaDataCloud project holds a specific task to work on the ingestion, validation, long-term storage and access of Flow Cytometer data. Also, OBIS is looking beyond biogeographic data. In the framework of OBIS-ENV-DATA an extension to the current Darwin Core data schema is defined to accommodate additional data types. This will allow for the management of sampling methodology, animal tracking and telemetry data, and environmental measurements such as nutrient concentrations, sediment characteristics or other abiotic parameters measured during sampling to characterize the environment from which biogeographic data was collected. This new structure enables the linkage of measurements or facts - quantitative or qualitative properties - to both sampling events and species occurrences.

2.1.5. Data policies

- The biological DMP for JERICO-NEXT is defined in line with the INSPIRE directive and consistency with other existing European infrastructures.
- We follow recommendations on methods for a free and open data access policy according to the biological data domain (see http://www.iobis.org/data/policy/).

2.1.6 Currently ongoing and planned activities

At the start of the project a data information survey was performed among the JRAP leaders to provide a clear view on the diversity of the biological data that are expected to be collected during the course of the project. In order to have a better idea on the timing of data delivery and both raw and processed data formats, this template was extended and sent out to the data providers and JRAP leaders in December 2016.

To streamline the integration exercise, the data management team needs the input from WP3 and WP4 (JRAP1&2). More specifically input on standard parameters, data formats and metadata (measurement, sensor).

Dataset metadata will be uploaded as described in section 2.1.3. This will make the datasets visible for the public under a Jerico-Next label on the EMODnet Biology webpage. The figure below shows an example. Eventually these metadata sheets will be linked to the location where the data will be accessible.

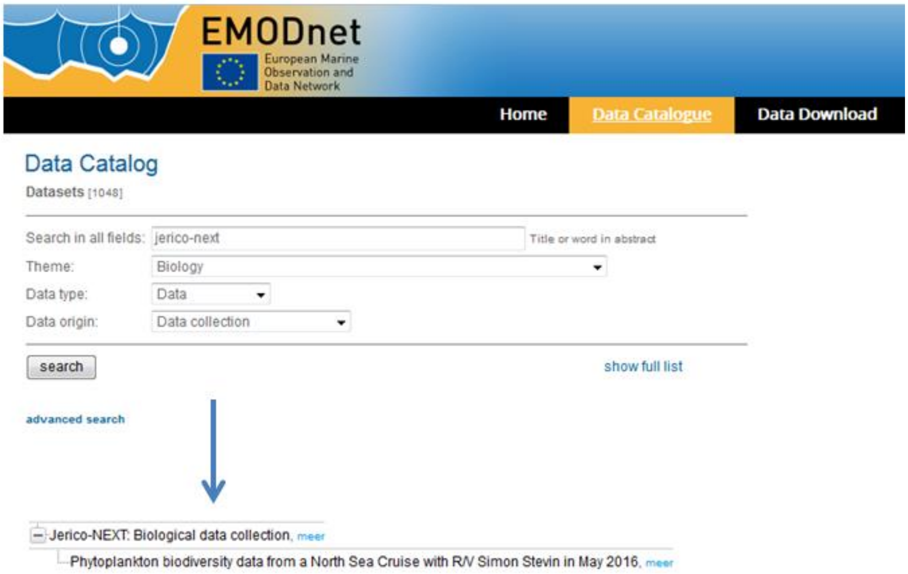

Figure 3: EMODnet Biology data catalogue search

Actions were carried out for the initial transfer of data by setting up a dedicated ftp server (jericonext@ftp.emodnet.eu).

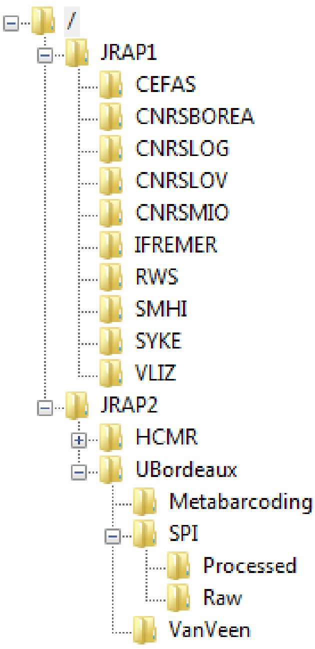

Figure 4: FTP server folder structure

Next to this short term FTP server, there is the possibility to store the data (raw or processed) for longer term on a Marine Data Archive.

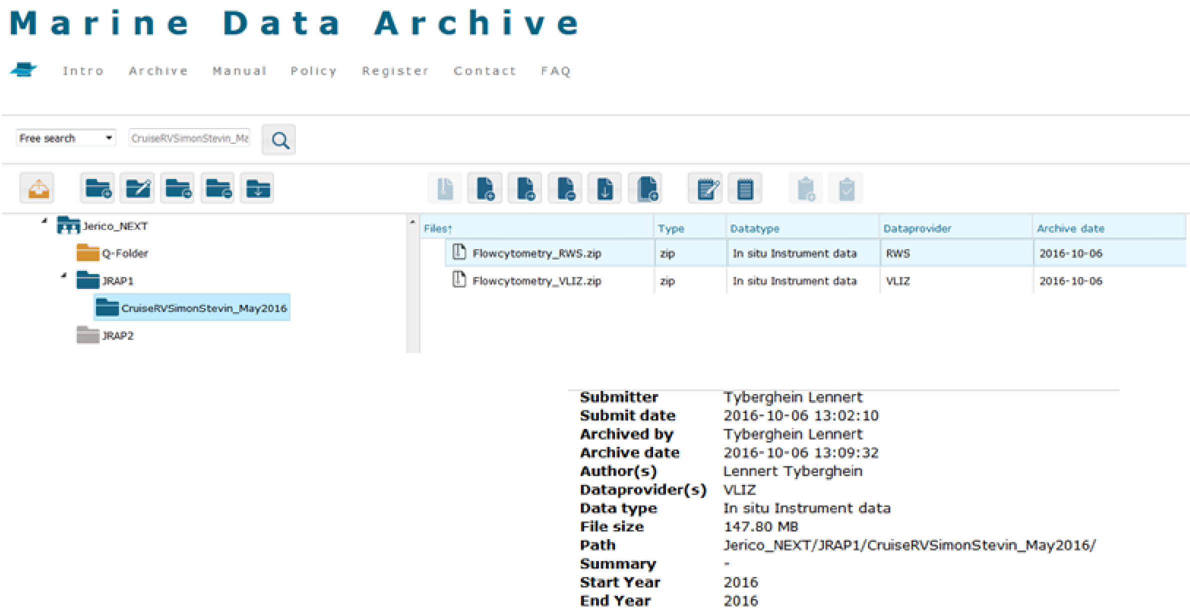

Figure 5: Screenshot of the Marine Data Archive.

In order to mobilize the scientists and data providers in this community we support the setup of dedicated user groups combining knowledge on specific sensors. Such group currently exists for the users of the Cytosense flow cytometer. Regular Webex meetings are set up to accomplish the above mentioned input for the data transfer and delivery.

The following timeline was agreed for the Jerico Next's Biological Data Integration procedure.

Time	Action
2017, Feb 1	Raw data examples delivered
2017, March	General Assembly
2017, April 1	Processed data examples delivered
2017, May 1	Description per sensor of what is raw & processed data and technical metadata
2017, June 1	Raw data of Year 1 delivered
2018, Feb 1	Raw data of Year 2 delivered
2018, May 1	JRAP1 workshop (present results)
2018, June 1	Processed data year 1&2 delivered
2018, Dec 1	Raw data Year 3 delivered
2019, Apr 1	Processed data Year 3 delivered
2019, May/June	Data release
2019, June	General Assembly
2019, Aug 1	Deliverable D5.5

Table 1: Task 5.2 timeline

2.2. Quality control procedures on biological data

In this section we describe procedures and methodologies that enable data collected through the project to comply with international standards regarding their quality and metadata. As such the description of QC procedures for biological data is an integral part of the data management plan.

Several levels of QC procedures can be distinguished. There are quality control checks and calibrations **at the level of the instrument or sampling device.** Although often ignored, these checks are a **prerequisite for good measurements**. At a second level, quality checks can be performed **on in situ measurements**. Further validations, **delayed mode QC** takes place once the data are stored in the data centers.

Here, we focus on the delayed mode QC steps that were developed by EurOBIS, in collaboration with the EMODnet and OBIS network. The biological data will be collected in WP3 and WP4 (JRAP1 and JRAP2). It consists of data that were collected by a mixed approach of both established and novel methodologies.

2.2.1. <u>Biogeographic data delayed mode QC procedures</u>

2.2.1.1. Pre-checks

A first step after the data has been harvested from a provider is to check if the data is comprehensible and consistent with the metadata provided (e.g. does the sampling protocol match what is expected, based on the assessed functional group, do the dates and coordinates fall within the range described by the metadata, etc.).

Other checks include:

- Checking that the required data fields are present and the values are possible.
- That all data fields contain the appropriate data.
- Database relational integrity for datasets which have Measurements or Facts measurements
- That abundances are provided for the datasets for which they were promised
- There are no 0 ('zero') values in the abundance and that the unit is known
- That when biomasses are provided, it is clear whether they are wet weight or dry weight
- When codes are provided for certain data (e.g. sex, lifestage, sampling gear,…), they are explained.
- Duplicate records. This check proved valuable not only to limit actual data duplication, but also to assess whether relevant sampling descriptors (different subsamples) or biotic measurements (e.g. life stages, size measurements) were omitted.

2.2.1.2. Automated procedures

After the upload of data into the EurOBIS database, it is possible to run a sequence of automated QC steps. For each record in the database, a Quality Control (QC) value is calculated based on which of these steps are passed.

The procedures can be put under five main categories (full detailed list: see annex 2):

Data format

- Checks if the required fields are filled in
- Checks of compliancy with the OBIS schema

Assessment of the completeness and validity of information

- Checks on detailed information (e.g. is the date realistic; is minimum depth smaller than maximum depth, …)

Taxonomic quality control
- Checks whether the taxon is listed in WoRMS, the World Register of Marine Species.

Geographic quality control
- Multiple checks on latitude, longitude (valid boundaries, not on land, not (0, 0))

Outlier analysis
- Checks for geographical outliers
- Checks for environmental outliers

For more details on the calculation of the QC values:
- Vandepitte, L.; Bosch, S.; Tyberghein, L.; Waumans, F.; Vanhoorne, B.; Hernandez, F.; De Clerck, O.; Mees, J. (2015). Fishing for data and sorting the catch: assessing the data quality, completeness and fitness for use of data in marine biogeographic databases. Database 2015: 14 pp.
- http://iobis.org/manual/lifewatchqc/
- http://iobis.org/manual/namematching/

2.2.1.3. Use of QC procedures

The OBIS website shows a limited set of metrics based on the automated QC procedure results. The data quality section on the right hand side shows for what percentage of records in the dataset the QC step was positive.


Overview

Records	16838
Taxa	931
Species	730
Node	EurOBIS
Provider	European Environment Agency
Resource	http://ipt.vliz.be/eurobis/resource?r=nsbs
Website	http://ipt.vliz.be/eurobis/resource?r=nsbs
Abstract	North Sea macrobenthos species density and abundance, measured on a regular grid of stations corresponding with ICES standard stations. All sampling was done in spring 1986.
Citation	Craeymeersh J., P. Kingston, E. Rachor, G. Duineveld, Carlo Heip, Edward Vanden Berghe, 1986: North Sea Benthos Survey.
License	Attribution-ShareAlike (CC BY-SA)
Taxon scope	Biota
Temporal scope	1985 to 1986
Geographical scope	Latitude 51.000 to 60.750; Longitude -3.000 to 9.000

Figure 6: Screenshot of an OBIS dataset search http://iobis.org/explore/#/dataset/586

2.2.1.4. Tools accessible for QC

The LifeWatch initiative, an E-Science European Infrastructure for Biodiversity and Ecosystem Research, translated a number of the above mentioned QC steps related to data format, taxonomy and geography to interactive, user-friendly web services (http://www.lifewatch.be/data-services). Using these services, data providers, data managers and users are able to make a general assessments of the quality, completeness and fitness for use of their own biogeographic data by simply uploading them to the LifeWatch portal and selecting the QC steps they want to run on their data. A detailed user guide and some use cases are provided.

2.2.2. Sensor data QC procedures

Within WP3 and WP4 a variety of sensor data (see annex 1) is generated. A large part of these sensors (incl. software) have large potential for monitoring purposes, but are currently still in a phase of development. As a consequence the data output is also highly variable and for most of the sensors there is not yet a consensus on protocols for data quality control. Quality control currently consists of expert validation of the data. It is the intention of this WP to document on these procedures and work towards standard output parameters for the different sensors. Once this is achieved and data is stored in a database, the above described delayed mode QC procedures (or parts) could be applied.

3. Executive Summary and Conclusions

This report provides information on the integration plan for biological data in the framework of JERICO-NEXT. It outlines the possibilities for the creation of operational links with EMODnet biology and OBIS. The procedures for integration of biological data into these systems will follow standard schemas and procedures. This document summarizes these procedures and provides links to additional information.

The detailed section on Quality Control procedures shows that a delayed mode of quality control is standard practice for biological data management. Data products derived from state of the art sensors should be standardized in the future. In that way a similar mode of QC could also be applied to these new kinds of data. Together with WP3 and WP4 (JRAP1 and JRAP2) this objective is put forward.

4. Annexes

Annex 1: Table listing data sources and derived parameters for WP3/WP4. It is indicated whether the derived data fits the common data schemas or new data schema need to be explored. Data archival is possible for all data.

JRAP	JRAP data sources (sensor or sampling technique)	(Derived) biological parameters	Fit for common data schemas	New data schemas to be explored	Data archival possible
1	(ft)-psicam	absorption, chlorophyll-a	no	yes	yes
1	frrf	Primary production	no	yes	yes
1	Cytosense flow cytometer	Total fluorescence per size class, Total biovolume per size class, characterisation of spectral groups	no	yes	yes
1	UVP5	Organisms per unit of volume	no	yes	yes
1	Imaging Flow cytobot	Phytoplankton biodiversity and abundance	yes	no	yes
1	fluorometer	Fluorescence parameters	no	yes	yes
1	fastcam	Phytoplankton biodiversity and abundance	yes	no	yes
1	phytoPAM	Fluorescence parameters	no	yes	yes
1	flowcam	Phytoplankton biodiversity and abundance	yes	no	yes
1	discrete water samples (microscopy)	Species abundance, Total abundance on higher taxonomic level	yes	no	yes
1	ferrybox data	Fluorescence parameters, Chlorophyll-a	no	yes	yes
2	SPI	OSI (sediment organism relationship) - BHQ (Benthic habitat quality quantification)	no	yes	yes
2	metabarcoding (dna data)	OTU abundance (Operational Taxonomic Units)	no	yes	yes
2	van veen grab	benthic species abundance	yes	no	yes

Annex 2:overview of all the QC-steps in the EurOBIS database, including the unique bit-sequence ($2^{(x-1)}$, with X=number of the QC flag) when the QC step is evaluated positively. The final column lists whether a QC step is also available to the users through the online web services. IQR=Interquartile range; MAD=Median absolute deviation; SSS=Sea surface salinity; SST=Sea surface temperature. (Vandepitte et al. 2015)

QC-number	Category	Question	bit-sequence, if answer is yes	available as online data service	Implemented in
2	Taxonomy	Is the taxon name matched to WoRMS?	2	Yes (taxon match)	EurOBIS + OBIS
3	Taxonomy	Is the taxon level lower than family?	4	Yes (taxon match)	EurOBIS + OBIS
4	Geography: lat/lon	Are the latitude/longitude values different from zero?	8	Yes (check OBIS format)	EurOBIS + OBIS
5	Geography: lat/lon	Are the latitude/longitude values within their possible boundaries?	16	Yes (check OBIS format)	EurOBIS + OBIS
6	Geography: lat/lon	Are the coordinates situated in sea or along the coastline (20km buffer)?	32	Yes (check OBIS format)	EurOBIS + OBIS
9	Geography: lat/lon	Are the coordinates situated in the expected geographic area (compare metadata)?	256	No, but visual check possible through separate data validation service	EurOBIS
18	Geography: depth	Is minimum depth ≤ maximum depth?	131072	Not yet available	EurOBIS + OBIS
19	Geography: depth	Is the sampling depth possible when compared to GEBCO depth map (incl. margin)?	262144	No, but depths per lat-lon can be requested through geographic web services	EurOBIS + OBIS
7	Completeness: date/time	Is the sampling year (start/end) completed and valid?	64	Yes (check OBIS format)	EurOBIS + OBIS
11	Completeness: date/time	Is the sampling date (year/month/day) (start/end) valid?	1024	Yes (check OBIS format)	EurOBIS + OBIS
12	Completeness: date/time	If a start and end date are given, is the start before the end?	2048	Yes (check OBIS format)	EurOBIS + OBIS
13	Completeness: date/time	If a sampling time is given, is this valid and is the time zone completed?	4096	Not yet available	EurOBIS + OBIS
14	Completeness: presence/abundance/biomass	Is the value of the field "ObservedIndividualCount" empty or > 0?	8192	Not yet available	EurOBIS + OBIS
15	Completeness: presence/abundance/biomass	Is the value of the field "Observedweight" empty or > 0?	16384	Not yet available	EurOBIS + OBIS
16	Completeness: presence/abundance/biomass	Is the field 'SampleSize' completed if the field 'ObservedIndividualCount' is > 0?	32768	Not yet available	EurOBIS + OBIS
1	(Eur)OBIS data format	Are the required fields from the OBIS Schema completed?	1	Yes (check OBIS format)	EurOBIS + OBIS
10	(Eur)OBIS data format	Is the 'Basis of Record' documented, and is an existing OBIS code used?	512	Yes (check OBIS format)	EurOBIS + OBIS
17	(Eur)OBIS data format	Is the value of the field "Sex" empty or is an existing OBIS code used?	65536	Not yet available	EurOBIS + OBIS
21	Outliers:environment	Is the observation within 6 MADs from the median depth of this taxon?	1048576	Not yet available	OBIS
22	Outliers:environment	Is the observation within 3 IQRs from the 1st & 3rd quartile depth of this taxon?	2097152	Not yet available	OBIS
23	Outliers:environment	Is the observation within 6 MADs from the median SSS of this taxon?	4194304	Not yet available	OBIS
24	Outliers:environment	Is the observation within 3 IQRs from the 1st & 3rd quartile SSS of this taxon?	8388608	Not yet available	OBIS
25	Outliers:environment	Is the observation within 6 MADs from the median SST of this taxon?	16777216	Not yet available	OBIS
26	Outliers:environment	Is the observation within 3 IQRs from the 1st & 3rd quartile SST of this taxon?	33554432	Not yet available	OBIS
27	Outliers:geography	Is the observation within 6 MADs from the distance to the centroid of this taxon?	67108864	Not yet available	OBIS
28	Outliers:geography	Is the observation within 3 IQRs from the 1st & 3rd quartile distance to the centroid of this taxon?	134217728	Not yet available	OBIS
29	Outliers:geography	Is the observation within 6 MADs from the distance to the centroid of this dataset?	268435456	Not yet available	OBIS
30	Outliers:geography	Is the observation within 3 IQRs from the 1st & 3rd quartile distance to the centroid of this dataset?	536870912	Not yet available	OBIS