Best practices and recommendations

for plankton imaging data management

Ensuring effective data flow towards international data infrastructures

Version 1

2022

Patricia Martin-Cabrera, Ruben Perez Perez, Jean-Olivier Irisson, Fabien Lombard, Klas O. Möller, Saskia Rühl, Veronique Creach, Markus Lindh, Lars Stemmann and Lennert Schepers.

Flanders Marine Institute, VLIZ (Belgium), Institut de la Mer de Villefranche, CNRS – Sorbonne Université (France), Centre for Environment, Fisheries and Aquaculture Science, CEFAS, (UK), Institute of Coastal Research, Helmholtz-Zentrum Geesthacht, Hereon Geesthacht, (Germany), Swedish Meteorological and Hydrological Institute, SMHI (Sweden).

This work was supported by the JERICO-S3 project under WP6- D6.4 and received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 871153.

Keywords: *Imaging data, plankton, best practices, Darwin Core, data management*

TABLE OF CONTENTS

1. Introduction 4

2. Plankton observations 5

2.1. Benchtop imaging instruments 6

2.2. *In situ* imaging 8

2.3. Image data processing 10

3. Data standards and best practices for imaging data management 10

3.1. OBIS-ENV-DATA format 11

3.2. How to populate the Event, Occurrence and EMoF tables for imaging data 12

3.2.1. Event table 12

3.2.2. Occurrence table 12

3.2.3. Extended Measurements or Facts (eMoF) table 14

4. Workflows: From instrument to EMODnet Biology 21

4.1. EcoTaxa workflow 22

4.2. EurOBIS harvest and EMODnet Biology publication 23

5. Quality control 23

6. Access & use 24

6.1. Dataset description 24

6.2. Dataset distribution and web services 25

6.3. Dataset citations 25

6.4. Dataset licence 25

7. Data Management Issues 26

8. Summary 28

9. Acknowledgments 28

10. Annexes 29

10.1. Imaging instruments and their BODC identifier. 29

10.2. Typical measurements derived from imaging and their BODC identifier. 29

References 30

Executive Summary

Plankton imaging instruments are increasingly used to record species occurrences, and they are also able to repeatedly measure ecological traits. However, due to the extensive variety of instruments and the different formats of the data output, there are currently no guidelines and best practices available to store all the relevant data and information in a standard format. Overcoming this challenge will allow for the integration and exchange of these datasets, enabling end users to analyse and visualise them more effectively.

To make these data as FAIR (Findable, Accessible, Interoperable, and Reusable) as possible and to share them with international biodiversity data portals, such as the European Marine Observation and Data Network (EMODnet Biology) and the international Ocean Biodiversity Information System (OBIS) Network, like EurOBIS (the European node of OBIS), best practices for the management of plankton imaging data are needed. Thus, the goal of this document is to provide recommendations to plankton imaging users on how to format their data following the OBIS-ENV-DATA format, a Darwin Core based approach to standardise biodiversity data, for submission to these international data portals. These best practices and recommendations are created by an expert working group in the framework of the JERICO-S3 project and by intensive interactions and feedback from the global marine plankton and OBIS community.

This document provides (1) an introduction of the current landscape of plankton imaging instruments and the processing of their images, (2) a description of the data standards and format used in biodiversity and guidelines on how to use these, (3) a workflow from instrument to EMODnet Biology, and (4) a discussion on the data management issues identified.

With the best practices presented here, it is possible to report a detailed taxonomic characterisation of plankton observations as well as quantitative information that is useful for ecological studies. This format allows biodiversity data portals to extend their scope beyond species occurrence data. Furthermore, proposing the use of more Darwin Core fields in this format, users now have the possibility to publish manually validated datasets, but also datasets produced by fully automated plankton identification workflows. The proposed data and file formats are simple and both human- and machine-readable to automatise workflows. This format will allow data generators to submit enriched plankton imaging datasets to the international biodiversity data portals, (Eur)OBIS and EMODnet Biology. We encourage plankton imaging data generators to implement these workflows into their pipelines, to share their data with the international data portals easily, enriching these databases with this valuable data.

1. Introduction

Imaging data can be defined as the qualitative and quantitative information derived from a collection of images (still images or videos). These data often include information on how sensors acquire the measurements and how the images are processed. Other valuable data from the images include qualitative and quantitative features, for example taxonomy and morphological measurements. Imaging systems are increasingly used in the marine domain. Over the last decade, plankton research has experienced an extensive development in automatic image acquisition for identifying and quantifying plankton species. These observations are promising and have several benefits. They generate huge amounts of data that can be acquired and processed very quickly. In addition, the workflow and resources required to collect and process data is highly cost-efficient compared to traditional methods such as microscopy. These improvements of imaging sensors, and the increased growth of the datasets generated, highlight the importance of adequate data management. Thus, establishing imaging best practices and recommendations, (semi) automated data flows and quality control procedures will promote the ability to make these datasets Findable, Accessible, Interoperable and Reusable (FAIR principles, Wilkinson et al, 2016) to ensure an operational use of ocean data for research.

The scope of this document is to propose recommendations and best practices for plankton imaging data management and data flows towards the European and international marine biodiversity data portals. The main portal is EurOBIS, the European Node of the international Ocean Biodiversity Information System (OBIS), and from there the data is shared with EMODnet Biology, the European Marine Observation and Data Network. Providing a standardised data format allows the submission of enriched data from the images acquired by plankton imaging instruments to these data portals.

Following the present biodiversity data standards and initiatives, the guidelines developed provide a set of recommendations on how to fill the OBIS-ENV-DATA format for plankton imaging data. This document describes the proposed format, including a dataflow that goes from instrument data to EcoTaxa, where the images are taxonomically annotated. It also proposes to use the export of data from EcoTaxa to the proposed format through an API that allows submitting the data to the biodiversity data portals easily. In addition, guidelines on how to format data following international standards are provided for users to follow, if taxonomic annotations are performed outside of EcoTaxa. Finally, we discuss how to aggregate the data to be able to report abundance or biomass of plankton species in a meaningful way.

The target users of this document are scientists generating data from plankton imaging instruments and data managers from National Oceanographic Data Centres (NODCs) handling these data. The geographical audience is Europe, however the workflows can also be adopted by an international audience, because data from the European data platforms can also flow to international platforms OBIS and the Global Biodiversity Information Facility (GBIF).

2. Plankton observations

Since the 1980s, a considerable amount of energy has been directed to produce prototypes of automated quantitative imaging instruments, some of which are now commercially available off-the-shelf, others even built by users. Instruments can be used in laboratory or at sea, on a benchtop or immersed in the water, but they all share some common principles:

- Marine particles and plankton either pass by or are placed in a known volume excited by a specific light source. Optical instruments make various optical measurements (e.g. fluorescence), while imaging instruments take an image, from which measurements are inferred. Both apply for the same object in the case of imaging flow cytometers.
- Images can be classified according to taxonomic or functional groups and living cells can be differentiated from aggregates and other non-living particles.
- Imaging software provides common particle characteristics: each object's size, shape and cross-sectional area can be determined, as well as the intensity of light coming from each pixel of the particle, identified thanks to its optical or image characteristics, producing a large amount of raw data. Sometimes, these data are used to provide statistics for a given group (e.g. flow cytometry) or for given sizes.
- Each optical/imaging technique also comes with its own size range limitation. However as a general characteristic, small particles are often too small to be imaged efficiently (too few pixels, signal below threshold or near noise level), while larger organisms are too scarce to be sampled quantitatively, if the volume analysed is too small, or too large to pass by the tubing of some instruments, resulting in a narrow size range for each instrument.
- Additionally, to obtain taxonomic information from optical or imaging methods, there is a need for a computer-assisted human expert to validate organisms based on their optical properties (e.g. "gating" in-flow cytometry) or on their image.

Lombard et al. (2019) provides a detailed review of existing plankton sampling technologies, from water samples to optical and imaging instruments, and targeted sizes varying from sub-µm (micrometres) to cm (centimetres), (Figure 1). In this section, we cover some of these technologies, dividing them into benchtop and *in-situ* instruments.

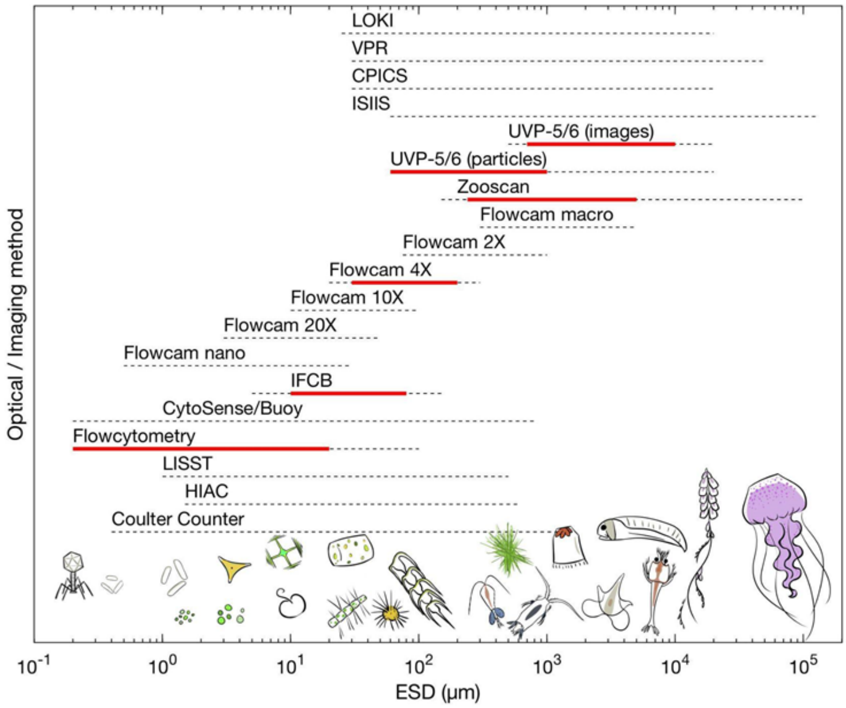

Figure 1: Comparison of the total size range of plankton (in equivalent spherical diameter; ESD) and optical and imaging analytical methods available. Dashed lines represent the total operational size range from commercial information while the red line represents the practical size range which is efficient to obtain quantitative information. Drawings by Justine Courboules. Redrawn from Lombard et al. (2019).

2.1. Benchtop imaging instruments

Benchtop imaging refers to devices that can be conveniently used on a workbench in the laboratory or research vessel. They require the collection of physical plankton samples. These can be obtained with a wide range of devices, such as plankton nets (e.g. WP2-net, multinet), bottles from CTD-Rosettes, or buckets from the surface. They also can be used as flow-through systems using pumps with a net system for collecting plankton samples. Benchtop instruments can be in the laboratory or on board of the research vessel. For example, the **ZooSCAN** (Gorsky et al., 2010) is a benchtop plankton scanner with custom lighting and a watertight scanning chamber into which the sample (liquid containing zooplankton organisms) is poured. The scanner makes a digital, high-resolution image of the sample of all objects above 200 µm ESD. Image resolution can be up to 4800 dpi (dots per inch) and each image is 14150 px (pixels) by 22640 px, containing hundreds to thousands of individual objects. The analysis is non-destructive since the liquid sample scanned can be recovered without damage through a drainage channel. The device also has built-in features making it possible to standardise the images of different ZooSCAN, to remote control the image generation, and to build a common image database.

Imaging Flow Cytometry (IFC) is a technique that combines features of flow cytometry (single-particle fluidics), optical characterization (fluorescent microscopy) and imaging of cells/colonies. IFC instruments can be used as benchtop and/or *in-situ*, and they differ in their approaches, outputs, and size range.

The **Imaging FlowCytobot** (IFCB) is a fully automated, submersible instrument that uses a combination of flow cytometric and video technology to capture high resolution images of suspended particles. Laser induced fluorescence and light scattering from individual particles are measured and used to trigger targeted image acquisition. IFCB generates high resolution (approximately 3.4 px/µm) images of suspended particles in the size range <10 to 150 µm. The instrument is very versatile, as it can be used *in-situ* connected with a flow-through system, for example a FerryBox, as well as for fixed deployments for continuous monitoring, enabling up to 6-9 months unattended deployments in the ocean, and also for benchtop analysis. It continuously samples at a rate of 15 ml (millilitre) of seawater per hour, generating on the order of 30000 high resolution images per hour, depending on the target population.

The **CytoSense** (and CytoSub which is a submersible version) is a portable, benchtop autonomous flow cytometer designed for phytoplankton species classification and analysis of filamentous algae. It can also be used *in situ* to reveal temporal and spatial phytoplankton variability. It can be remotely controlled and has been specifically designed to record the optical pulse shapes of suspended particles between <1 and 800 µm in diameter and up to 4 mm (millimetres) in length (for chain-forming cells recording) in relatively large volumes of water (several cm cubed per sample). The instrument combines high sensitivity with an extremely wide particle size range (from sub-µm up to 1.5 mm in diameter) and acquires multiple data points per particle, which distinguishes the CytoSense from conventional flow cytometers. The sample intake speed ranges from 0.07-17 µl/s (microlitres per second), allowing high particle loads (thousands of particles per second) as well as very low concentrations. The CytoSense has a modular design, with various upgrades, including additional lasers, optional cameras for imaging of particles and a widened flow cell.

The **FlowCam** VS series is a benchtop automated imaging-in-flow instrument that generates high-resolution digital images for measuring size and shape of microscopic particles. The sample introduced in the system is attracted by a peristaltic or a syringe pump into a flow cell (or flow chamber) with known dimensions, located in front of a microscope objective which is connected to a camera video. It uses a similar imaging principle as IFCB but lacks the hydrodynamic focusing provided by the sheath flow. Images are acquired either continuously (auto trigger mode) or after the detection of a fluorescent (chlorophyll *a*) particle. The instrument can measure particles between 2 µm and 2 mm. It has a flow rate between 0.005 ml/minute and 250 ml/minute, depending upon magnification, flow cell depth, camera frame rate, efficiency desired, etc. FlowCam VS is available in four models, from the imaging-only VS-I (i.e. without excitation wavelength or fluorescence emission wavelengths) to the top-of-the-line VS-IV with two channels of fluorescence measurement and scatter triggering capabilities. It can produce either 8-bit grayscale (monochrome camera) or 24-bit colour (colour camera) images, depending on the model.

The **ZooCAM** (Colas et al., 2018) is an in-flow system for on-board imaging of large volume samples of preserved and living metazooplankton (i.e. multicellular zooplankton) and fish eggs >300 µm ESD. The ZooCAM features a fluidic module and an optical/imaging module. The sample is mixed with filtered seawater in the fluidic module, which is connected to a high-volume peristaltic pump. The pump drives the seawater and the particles through the tubing to a flow cell where they are imaged. The flow cell is mounted between the camera and the illumination system. The images are captured by a 1280 x 1024 pixel black and white USB 3.0 CCD camera (Thorlabs), on which a telecentric 0.5x lens (Edmund optics) is mounted. The pump flow speed can be manually adjusted between 0.28 l/min (litres per minute) and 1.7 l/min. It uses an imaging principle similar to that of the FlowCam-Macro.

The **PlanktoScope** uses an imaging principle similar to that of the FlowCam but is a low cost, open-source imaging instrument designed for citizen science (Pollina et al., 2020). It is a flow microscope capable of autonomously imaging 1.7 ml/minute with a 1.5 µm resolution, pumping samples in a flow chamber. The computing module is made of the latest Raspberry Pi 4 (4GB of LPDDR4 SDRAM) coupled with its Pi Camera v2.1 - 8 Mpx (Megapixels). The optic is simplified using two reversed M12 lenses, the tube lens is fixed when the objective lens can be swapped offering a variety of optical configuration.

2.2. *In situ* imaging

In situ imaging allows the study of plankton directly at sea without the need of collecting water samples, nets or pumping water. These instruments can be deployed on fixed platforms (e.g. cabled observatories), in free drifting floats, and towed or deployed vertically from the ship. Compared to the traditional techniques, *in-situ* imaging is considered less destructive because when collecting samples, these can be damaged or modified if treated with fixatives.

The **Ichthyoplankton Imaging System** (ISIIS, Cowen and Guigand, 2008) is an underwater imaging system for capturing *in situ*, real time images at a fine spatial and temporal resolution. It captures a wide taxonomic range of mesozooplankton, such as fish larvae and fragile gelatinous organisms, and with lower resolution, large protists, and cyanobacteria. The organisms are imaged as they swim or flow in between the two pods of the instrument, which is fitted with a camera and illumination system. On towed sleds, they use a line-scan camera creating one single continuous image representing a real slice of the ocean. Moreover, they can be fitted with a classic area-scan camera if a system is to be used still (underwater monitoring station) or do slow vertical profiles. ISIIS is capable of imaging a maximum of 162 litre) of water per second (when flying at 5 knots) with a pixel resolution of 70 µm, imaging particles from 1 mm to 13 cm in size.

The **Continuous Plankton Imaging and Classification System** (CPICS) is an underwater imaging system that can be deployed on a CTD-Rosette and free-drifting or anchored instrument platforms, both static and vertically profiling. Deployments can be in a self-contained format, relying on internal batteries and file storage, or with a real-time power and/or data link to a vessel or the shore. The CPICS uses darkfield illumination to capture high-resolution colour images between 20X to 0.16X (X=magnifications), showing features ranging from 0.04 mm to 12 mm. It has a colour resolution of 24-bits, an image resolution of 6 Mpx (2736 x 2192) and a maximum frame rate of 10 fps (frames per second), with a depth rating of between 1000 m and 6000 m depending on the model used. Synchronisation of light output from the LED ring light system and exposure duration < 10 µs (microseconds), facilitates the capturing of images without motion artefacts, even at current or profiling speeds higher than 5 m/s (metres per second). The sampling volume is adjustable, depending on the area and focus threshold settings chosen by the user to optimise image quality and quantity in different levels of turbidity. Because of its open-flow approach to water sampling, delicate structures of plankton and particles remain intact, as do predator-prey interactions.

The **Video Plankton Recorder** (VPR, Davis et al., 2005) is an underwater video microscope system allowing image generation of plankton and particulate matter particles from 50 µm to a few cm in size. It can be vertically deployed or attached to towed bodies, Remotely Operated Vehicles (ROVs), Autonomous Underwater Vehicles (AUV) and moorings at depth ratings of 350 - 1000 m. A video camera of 1 Mpx (10 Bit black and white or colour) mounted in one of the arms focuses on a point midway between the two arms and a strobe on the other arm illuminates the imaged volume in between. It images objects with 15-25 fps in the water column with a small volume of seawater (1 ml to 350 ml depending on calibration) based on dark-field-illumination. Images are sent in real time on board or shore via a fibre optic tow cable while the Digital Autonomous Video Plankton Recorder (DAVPR) is fully self-contained.

The **Underwater Vision Profiler 5 HD** (UVP5, Picheral et al., 2010) is designed to image >100 µm particles and >500 µm zooplankton, with a 4 Mpx camera, a field of view of approximately 180x180 mm² about 200 mm in front of the camera. It can be operated as a stand-alone system, or attached to a CTD-Rosette, ROV, AUV or mooring. The UVP5 acquires only in-focus images in a volume of water delimited by a light beam issued from red light-emitting diodes (LEDs) in 100 µm flashes. The typical light beam illuminates an area of 4x20 cm which gives a sampling volume of 1 litre per image. The imaging rate is 20 images per second. The **UVP6-LP** (Picheral et al., 2021) is a miniaturised and low power version of the UVP5, designed for low speed, and limited space, to be deployed in ARGO floats, moorings, AUVs or gliders. Unlike the UVP5 the UVP6-LP cannot be used on a CTD-Rosette due to its 1.3 Hz (hertz) low acquisition frequency and 500 µm flashes. It acquires only in-focus images in a volume of water delimited by a single red flashing light illuminating a volume of 0.65 l. Optionally a 0.1% accuracy pressure sensor can be added. The resolution is 5 Mpx/0.73 µm, with a field of view 180 x 151 mm, and maximum image frequency 1.3 Hz. The operational depth for UVP5 and UVP6 is 0 to 6000 m.

The **Lightframe On-sight Keyspecies Investigation** (LOKI, Schulz et al., 2010) is an underwater camera system designed for vertical hauls for continuous, *in-situ* imaging of zooplankton, using a flow-through chamber with an upstream plankton net. It sends data every second, providing a vertical resolution of zooplankton and environmental variables of approximately 30-40 cm. LOKI operates an industrial camera with up to 6 Mpx at 15 µ*s* shutter time, combined with a tailored high power LED flash unit to image a volume of approximately 20x20x5 mm³ in a flow through chamber.

There are also submersible digital holographic particle imaging instruments that allow for 3-D reconstruction of images. They take images at very short shutter times, scanning for a relatively large volume. For example the **LISST-HOLO** measures large, complex flocs, plankton, and other particles in water. It contains a red (658nm) laser that emits collimated light into the sample volume. The light is scattered by suspended particles. The scattered light then interferes with the unscattered portion of the beam. The resulting interference pattern is captured by an onboard camera. The image captured by the camera is known as a hologram. The hologram can be digitally reconstructed to produce an in-focus picture of all the particles in the sample volume, from which particles size, shape, and position can be extracted.

2.3. Image data processing

Image data processing includes (1) sorting and cropping the images to obtain images with individual specimens, (2) image classification and annotation to assign a taxon performed with an automated classifications and (3) feature extraction to determine, for example, morphological features (e.g. Equivalent Spherical Diameter or ESD), performed by feature extraction software. In some instances, the first step is achieved automatically by the imaging instrument through internal detection and cropping algorithms, before images are even saved (e.g. in the case of CPICS).

Due to the massive number of images, the identification of plankton and other particles derived from imaging are performed with a software. Identifications made by a software, also called classifications, are commonly performed automatically using a machine-learning classifier. These classifications can be manually validated by an expert. Image processing software can be specific for certain instruments (e.g. the Deep Learning Image Classification Environment (DICE), which is a commercially available image classifier based on CEVA Deep Neural Network techniques, designed to accompany the CPICS system), others can be more general and used for several instruments (e.g. openly available Python and/or MATLAB scripts such as the YOLO classifier (Redmon et al. 2016) or the morphological features toolbox).

Plankton imaging data often includes measurements derived from the images. These measurements provide for example information about the size of the object from which biotic measurements such as cell or body size and biovolume (as a proxy of biomass) can be inferred.

3. Data standards and best practices for imaging data management

Data standards ensure interoperability between and within repositories, facilitating the integration, sharing, discovery, and long-term reuse of a dataset (FAIR principles). Interoperability is one shared goal among the plankton imaging community, therefore controlled vocabularies are essential to achieve this. Both EurOBIS and EMODnet Biology use the OBIS-ENV-DATA format (De Pooter et al., 2017) which relies on international standard in biodiversity informatics, the Darwin Core (DwC) standard for biodiversity terms (Wieczorek et al., 2012). In addition, it uses a number of marine specific controlled and standardised vocabulary terms in order to make data more interoperable, such as the BODC vocabularies to standardise parameters that are not covered by DwC, WoRMS the authoritative taxonomic list and catalogue of marine species names and the Marine Regions standard for marine georeferenced place names and areas.

The **DwC standard** includes **DwC terms** that are used to facilitate the sharing of information about biological diversity providing identifiers, labels, and definitions. To develop these best practices, we examined in detail the DwC terms used in the OBIS-ENV-DATA format, to follow this as much as possible and adapt it to our needs (see section 5.1).

BODC vocabularies are controlled vocabularies used in oceanographic data, managed and hosted by the British Oceanographic Data Centre (BODC) by means of the NERC Vocabulary Server. The BODC vocabularies present several collections of standardised terms to be able to understand the meaning of each term, and to enable machine to machine interoperability. The OBIS-ENV-DATA format uses these in one of the tables to store a machine-readable label (URI or URL).

New controlled vocabularies that are essential to provide provenance information of imaging datasets were identified. The vocabularies for eleven imaging instruments were created and stored in the L22 collection of the SeaVoX device catalogue, whose purpose is to define and describe instruments used for measurements at sea. Additionally, a list of vocabularies was created in P01 and P06 collections for annotating individual measurements derived from the images, that allow the computation of final concentrations. In section 5.2 of this document there is a detailed explanation on how to include these vocabularies, and in Annex 10.1 and 10.2, there is a complete list of the vocabularies created.

WoRMS provides an authoritative and comprehensive list of names of marine organisms, including information on synonymy. All synonyms are included in the register, allowing to standardise the names used in different datasets. To submit data to EMODnet Biology, a reference to the Life Sciences Identifier (LSID) assigned to the taxon in Worms is required (urn:lsid:marinespecies.org:taxname:1080). Additionally, names can be easily matched with their correct LSID using the Taxon Match tool of WoRMS. See table 2.2 for a practical example that includes the LSID from WoRMS. **The Marine Regions** database provides standardised marine georeferenced place names and areas. We encourage the use of the Marine Regions Gazetteer to define the most relevant sea area for the geographic coverage of the dataset. This information is included in the metadata of the dataset. It is a requirement for submissions to EurOBIS/EMODnet Biology, serving as a geographic quality control during the process of harvesting the dataset. For example if the geographic coverage in the metadata is “North Sea” (http://marineregions.org/mrgid/2350), but a number of data points in the dataset are outside of this area, this may indicate errors, and should be checked with the data provider.

3.1. OBIS-ENV-DATA format

Big efforts were made towards the development and adoption of the OBIS-ENV-DATA format to be able to ingest additional information related to the sampling activity in an interoperable manner in EurOBIS and EMODnet Biology (De Pooter et al., 2017). This format consists of an addition of the Darwin Core (DwC) Extended Measurement Or Facts or eMoF extension to the DwC Event core. The eMoF is linked to the occurrence table and allows storing biotic, abiotic and sampling measurements and facts related to the occurrence. An important aspect of this extension is that eMoF allows us to include standardised terms and controlled vocabularies. In plankton imaging data, it is crucial to describe the sample processing protocol to be able to cross calibrate the information originated from the different imaging instruments. However, the current structure of the OBIS-ENV-DATA format does not include sufficient information for imaging data. In this work, we aimed to include additional information in this format, including more details about the identification and to report quantitative information, with the objective to increase the transparency, provenance, and usefulness of this data (e.g. investigating ecosystem functioning and determining temporal and/or spatial distribution patterns).

3.2. How to populate the Event, Occurrence and EMoF tables for imaging data

The data structure consists of three flat tables: the **Event table**, that stores sample or observation information (time, location, depth, event hierarchy), the **Occurrence table**, that stores details of the sample or observation (e.g. identification details and taxonomy) and the **Extended Measurements or Facts** (eMoF) table, that allows storing additional biological and abiotic information from the events and their occurrences.

The three tables are related to each other by using the fields: *eventID* and *occurrenceID* (Figure 2). The *eventID* links the eMoF table to the Event Core. The *occurrenceID* is used to link biotic measurements in the eMoF table with the Occurrence table. The column names of the three tables must follow the DwC terminology.

The following sections include detailed information (based on the OBIS-ENV-DATA format) on each of the fields of these tables, indicating (1) required or optional fields, (2) definition and content expected, and (3) example for imaging data. An explanation on how these fields are populated is available in the OBIS manual, and an overview of this format can be found here https://www.eurobis.org/data_formats. The following sections highlight recommendations and additions of existing DwC terms to the OBIS-ENV-DATA format for plankton imaging datasets with a detailed explanation on how to populate the fields specifically for imaging datasets.

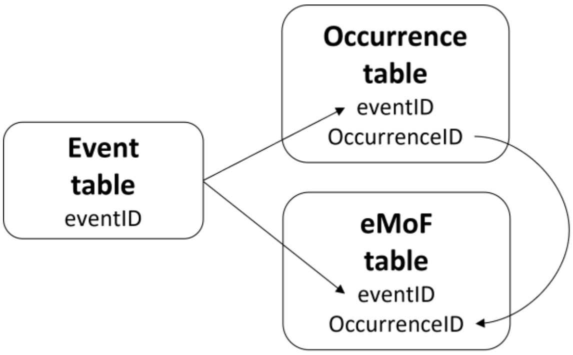

Figure 2: OBIS-ENV-DATA structure showing how the tables are linked to each other.

3.2.1. Event table

This table only refers to the event information with no specifications for imaging data needed. See the EurOBIS and EMODnet Biology guidelines template for detailed information about the fields on this table and guidance on how to populate it.

3.2.2. Occurrence table

Only taxonomical identified organisms should be stored in the table. In this section we mention the fields that are crucial for imagery datasets, with the DwC definition and identifier, and our recommendation on how to populate them, followed by a practical example (Table 1).

- ***basisOfRecord:*** Required in EurOBIS
  - DwC definition: The specific nature of the data record.
  - Identifier: http://rs.tdwg.org/dwc/terms/basisOfRecord

Recommended best practice is to always use the term of MachineObservation for imaging datasets derived from imaging instruments.

- ***identifiedBy:*** Necessary for imaging data if data has been validated
  - DwC definition: A list (concatenated and separated) of names of people, groups, or organisations who assigned the Taxon to the subject.
  - Identifier: http://rs.tdwg.org/dwc/terms/identifiedBy

If the identification has been validated by human(s), recommended best practice is to add the name(s) of the persons involved in verifying the automatic identification made by the algorithm/software. This field is useful to retrieve imaging datasets from EurOBIS where identifications of organisms have been validated by human, if:

basisOfRecord=MachineObservation + *identifiedBy* = is filled= Validated Imaging dataset.

basisOfRecord=MachineObservation + *identifiedBy* = is not filled= Non-Validated Imaging dataset.

- ***identificationVerificationStatus:*** Necessary for imaging data
  - DwC definition: A categorical indicator of the extent to which the taxonomic identification has been verified to be correct.
  - Identifier: http://rs.tdwg.org/dwc/terms/identificationVerificationStatus

This field holds information about the degree of uncertainty of the identification. Recommended best practice is to use the following categories:

- PredictedByMachine: for identifications generated by an algorithm and not validated by human.
- ValidatedByHuman: for identifications generated by an algorithm and verified to be correct by a human, this is also referred as validated data.

It is crucial to check the field *identificationVerificationStatus,* to be certain of the correctness of the identification. This field is useful to select validated or non-validated*,* if:

basisOfRecord=MachineObservation + *identificationVerificationStatus*= PredictedByMachine, users can not be fully confident on the identification of the records.

basisOfRecord=MachineObservation + *identificationVerificationStatus*= ValidatedByHuman, users can be confident of the identification of the records.

- ***identificationReferences:*** Necessary for imaging if data has not been validated
  - DwC definition: A list (concatenated and separated) of references (publication, global unique identifier, URI) used in the Identification.
  - Identifier: http://rs.tdwg.org/dwc/terms/identificationReferences

Recommended best practice is to add the citation (including the version) of the software, and/or the algorithm that did the identification (e.g. Plankton Identifier).

- ***associatedMedia:*** New to EurOBIS database (optional). Proposed in Neeley et al. (2021) for submissions to OBIS.
    - DwC definition: A list (concatenated and separated) of identifiers (publication, global unique identifier, URI) of media associated with the Occurrence
    - Identifier: http://rs.tdwg.org/dwc/terms/associatedMedia

Recommended best practice is to provide a unique persistent URL pointing to the landing page that contains the annotated images from which the occurrences are derived. This can be also a link to a .zip file or for example in EcoTaxa, a link to the project where the images are hosted (e.g. https://ecotaxa.obs-vlfr.fr/prj/15).

Table 1: A practical example of how to populate the Occurrence table (not all required fields are presented here).

eventID	*occurrenceID*	*basisOf Record*	*identified By*	*identificationVerification Status*	*identification References*	*associated Media*	*scientificName*
StationVisit:1_Sample:1_Subsample:1	StationVisit:1_Sample:1_Subsample:1_Taxon:106422_1	MachineObservation		PredictedByMachine	Plankton Identifier V1.3.4	https://ecotaxa.obs-vlfr.fr/prj/15	Oithonidae
StationVisit:1_Sample:1_Subsample:2	StationVisit:1_Sample:1_Subsample:1_Taxon:346029_1	MachineObservation	Patricia Cabrera	ValidatedByHuman	Plankton Identifier V1.3.4	https://ecotaxa.obs-vlfr.fr/prj/15	Acartia nana
StationVisit:1_Sample:1_Subsample:3	StationVisit:1_Sample:1_Subsample:1_Taxon:104079_1	MachineObservation		PredictedByMachine	Plankton Identifier V1.3.4	https://ecotaxa.obs-vlfr.fr/prj/15	Calanidae
StationVisit:1_Sample:1_Subsample:4	StationVisit:1_Sample:1_Subsample:1_Taxon:104079_2	MachineObservation		PredictedByMachine	Plankton Identifier V1.3.4	https://ecotaxa.obs-vlfr.fr/prj/15	Calanidae

3.2.3. Extended Measurements or Facts (eMoF) table

The eMoF table contains the fields “eventID” and “occurrenceID” to be able to link the biological and abiotic information in this table to their events and their occurrences. The measurements of facts are stored using the fields *measurementType*, *measurementValue* and *measurementUnit*, which are free text fields and human readable. Whereas the column names ending in ‘ID’: *measurementTypeID, measurementValueID* and *measurementUnitID,* are populated with controlled vocabularies, using the Unique Resource Identifiers (URIs) from the BODC Vocabulary Server (e.g. http://vocab.nerc.ac.uk/collection/P01/current/SDBIOL01/). It is highly recommended to always populate the *measurement(Type/Value/Unit)ID* fields, when there are corresponding URIs (exact matches only), to allow standardisation of these data and to improve data interoperability. Vocabularies can be searched per collection via the BODC vocabulary search, or if they do not exist they can be requested on their dedicated GitHub repository at https://github.com/nvs-vocabs (see: Request for new terms) . The following are the most relevant collections per field used in OBIS-ENV-DATA format, and relevant for imaging:

- ***measurementTypeID:*** Use collections OBIS sampling instruments and methods attributes Q01, and BODC Parameter Usage Vocabulary P01,
- ***measurementValueID:*** Use collections SeaVoX Device Catalogue L22 and Biological entity life stage terms S11,
- ***measurementUnitID:*** Use collection Approved data storage units P06

For imaging datasets, the information stored in the eMoF table concerns data about the sampling protocol. This information is useful in case users want to retrace how the final concentrations were calculated. Other information stored in the eMoF is the data derived from the images, (e.g. morphological measurements), and environmental variables (e.g. temperature and salinity). To report sampling acquisition data, it is important to distinguish between benchtop and *in-situ* imaging instruments, because both differ in the way samples are collected and analysed, and therefore the fields used will also be different. For benchtop imaging instruments, samples are collected at sea by traditional methods (e.g. nets) and processed with the imaging instrument. For this, we recommend populating the eMoF as indicated in Figure 3 and Tables 2.1-2.2.

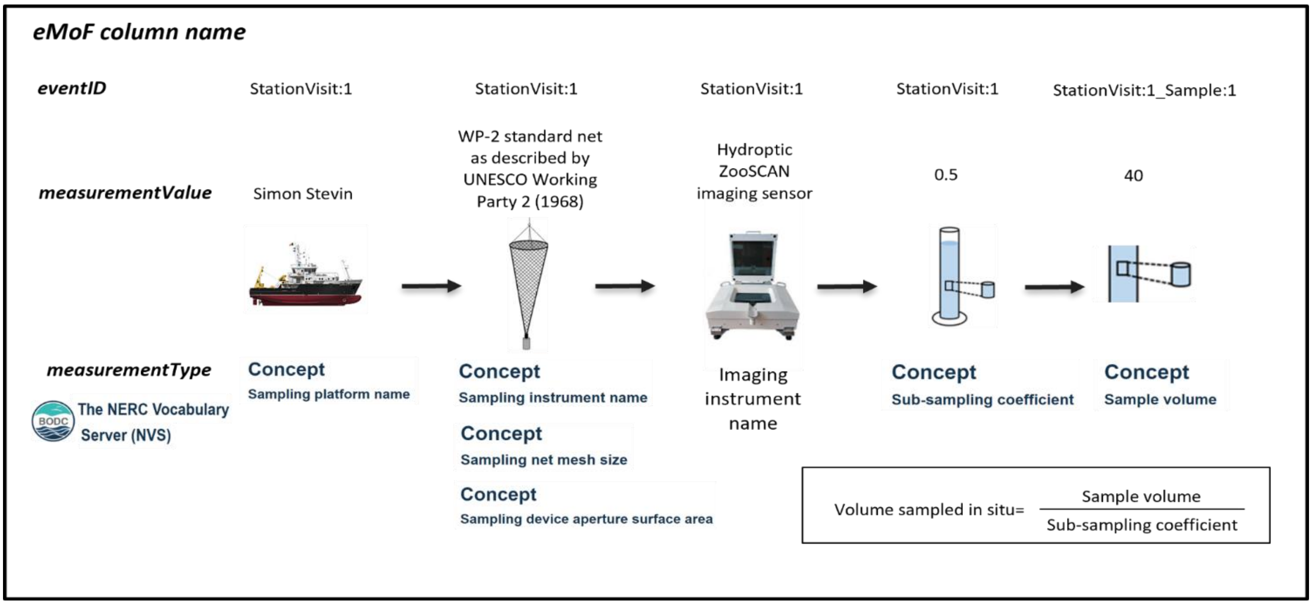

Figure 3: Schematic representation of sample collection for the ZooSCAN and how to populate this information on the eMoF table with the corresponding BODC vocabularies. In this example, if for example a Motoda box is used to split samples in halves, then "*Subsampling coefficient*"=0.5, and the final imaged volume *"Sample volume"*= 40 litre. Then, the volume sampled *in situ*= 40/0.5= 80 litre.

Table 2.1: Event table following the example shown above.

eventID	*type*	*parentEventID*	eventDate	decimalLongitude	decimalLatitute
StationVisit:1	station visit		2022-01-01	3	52
StationVisit:1_Sample:1	sample	*StationVisit:1*			

Table 2.2: eMoF table following the example shown above.

eventID	*measurementType*	*measurementTypeID*	*measurementValue*	*measurementValueID*	*measurementUnit*	*measurementUnitID*
StationVisit:1	Sampling platform name	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100001/	Simon Stevin	http://vocab.nerc.ac.uk/collection/C17/current/11SS/	not applicable	https://vocab.nerc.ac.uk/collection/P06/current/XXXX/
StationVisit:1	Sampling instrument name	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100002/	WP-2 standard net as described by UNESCO Working Party 2 (1968)	http://vocab.nerc.ac.uk/collection/L22/current/NETT0168/	not applicable	https://vocab.nerc.ac.uk/collection/P06/current/XXXX/
StationVisit:1	Sampling net mesh size	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100015/	200		Micrometres (microns)	http://vocab.nerc.ac.uk/collection/P06/current/UMIC/
StationVisit:1	Sampling device aperture surface area	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100017/	0.25		Square metres	http://vocab.nerc.ac.uk/collection/P06/current/UMSQ/
StationVisit:1	Imaging instrument name	BODC Term requested	Hydroptic ZooSCAN imaging sensor	http://vocab.nerc.ac.uk/collection/L22/current/TOOL1581/	not applicable	https://vocab.nerc.ac.uk/collection/P06/current/XXXX/
StationVisit:1	Subsampling coefficient	http://vocab.nerc.ac.uk/collection/P01/current/SSAMPC01/	0.5		not applicable	https://vocab.nerc.ac.uk/collection/P06/current/XXXX/
StationVisit:1_Sample:1	Sample volume	http://vocab.nerc.ac.uk/collection/P01/current/VOLXXXXX/	40		Millilitres	http://vocab.nerc.ac.uk/collection/P06/current/VVML/

In this practical example, samples are collected aboard the Simon Steven research vessel, and this information is reported using the field "*Sampling platform name"*. The field "*Sampling instrument name"* is used to indicate the name of the gear or instrument used to collect the sample, the WP2 net. Details about the net are stored using the BODC terms "*Sampling net mesh size"* and "*Sampling device aperture surface area"*. The instrument acquiring the images is populated in the field "Imaging instrument name". When samples are sub-sampled, the coefficient that indicates the proportion of the sample that is represented in the sub-sample is indicated in the field "*Subsampling coefficient".* Adding the "*Subsampling coefficient"* is relevant because if a concentration is computed from half a sample (Subsampling coefficient=0.5) it is more trustworthy than when the Subsampling coefficient=0.00001. This is important for converting from volume analysed by the imaging instrument (*"Sample volume")* to total volume of seawater sampled.

To indicate the relation between events and sub-events (in this case, samples and sub-samples) we use event hierarchy. We link events and their sub-events using the parentEventID and the type fields. Bear in mind that the use of the type field is currently being discussed and may be replaced in the near future by eventType, see here.

If the method used to divide the sample into subsamples is complex and requires additional important information this can be stored using the BODC term *Subsampling protocol*, in the eMoF.

Reporting occurrences and concentrations/biovolumes

Quantitative information is reported in the eMoF table. Concentration or biovolume is stored using the BODC terms: "Abundance of biological entity specified elsewhere per unit volume of the water body" or "Biovolume of biological entity specified elsewhere per unit volume of the water". This value is computed per sample and taxon.

The reason to report concentrations at a higher taxon is because when one or a few organisms from a specific taxon were identified haphazardly but not thoroughly looked for in all samples, we recommend reporting the occurrence, but not the concentration, because these may be underestimated (not all organisms from that taxon were found). In the example in tables 3.1-3.2, there are 5 occurrences, from which 4 correspond to the occurrences at the most detailed taxonomic level, and the fifth occurrence correspond to the aggregated 4 occurrences at a coarser taxonomic level (Copepoda) at which the operators are confident the counting was exhaustive to report concentration. This record is treated as a new occurrence and will have a unique new *occurrenceID.*

It is also possible to report the individual measurements of each of these occurrences. This information is useful if users would like to include size trait information in the dataset. These measurements (e.g. length or width of the organisms in the image or ESD) are also reported in the eMoF table as shown in Table 3.2.

Tables 3.1-3.2: Reporting occurrences and concentrations/biovolumes

Table 3.1: Occurrence table.

eventID	*ocurrenceID*	*scientificName*	*scientificNameID*
StationVisit:1_Sample:1_Subsample:1	StationVisit:1_Sample:1_Subsample:1_Taxon:106422_1	Oithonidae	urn:lsid:marinespecies.org:taxname:106422
StationVisit:1_Sample:1_Subsample:2	StationVisit:1_Sample:1_Subsample:2_Taxon:346029_1	Acartia nana	urn:lsid:marinespecies.org:taxname:346029
StationVisit:1_Sample:1_Subsample:3	StationVisit:1_Sample:1_Subsample:3_Taxon:104079_1	Calanidae	urn:lsid:marinespecies.org:taxname:104079
StationVisit:1_Sample:1_Subsample:4	StationVisit:1_Sample:1_Subsample:4_Taxon:104079_2	Calanidae	urn:lsid:marinespecies.org:taxname:104079
StationVisit:1_Sample:1	StationVisit:1_Sample:1_Taxon:1080	Copepoda	urn:lsid:marinespecies.org:taxname:1080

Find here a full example with the examples above.

Grouping occurrences

Two situations can happen for this grouped record regarding the status of the identification:

- When all the identifications of the organisms grouped in the same taxon have the same identificationVerificationStatus (e.g. ValidatedByHuman), only one unique *occurrenceID* is needed. The summed concentration of all organisms is reported in the eMoF table in the field “Abundance of biological entity specified elsewhere per unit volume of the water body”.
- When the grouped taxon is comprised of organisms that contain more than one identificationVerificationStatus (e.g. ValidatedByHuman and PredictedByMachine), 2 unique *occurrenceIDs*, are needed. These two will have the same taxon, but a different *identificationVerificationStatus* and their corresponding summed concentration from the organisms with the given identificationVerificationStatus, as well reported in the eMoF table in the field “Abundance of biological entity specified elsewhere per unit volume of the water body”.

This division of occurrences can also be done for the combination of data based on other parameters with for example, different life stages or size classes. Specifying the status of the identification allows users to decide what data to use. If users prefer to choose only human validated data, there is a risk of underestimating concentrations of a certain taxon (except if the dataset contains only human validated data). If users prefer to choose the entire dataset with validated and non-validated data, the concentrations of all organisms in the same taxon can be summed, risking errors on the identification made by a machine.

Table 3.2: eMoF table with only occurrence related measurements or facts.

eventID	*ocurrenceID*	*measurementType*	*measurementTypeID*	*measurementValue*	*measurementUnit*	*measurementUnitID*
StationVisit:1_Sample:1	StationVisit:1_Sample:1_Taxon:1080	Abundance of biological entity specified elsewhere per unit volume of the water body	http://vocab.nerc.ac.uk/collection/P01/current/SDBIOL01/	0.8743	Number per cubic metre	http://vocab.nerc.ac.uk/collection/P06/current/UPMM/
StationVisit:1_Sample:1_Subsample:1	StationVisit:1_Sample:1_Subsample:1_Taxon:106422_1	Length (expressed as pixels) of biological entity specified elsewhere	http://vocab.nerc.ac.uk/collection/P01/current/LGPIXEL1/	5	Pixels	http://vocab.nerc.ac.uk/collection/P06/current/PIXY/
StationVisit:1_Sample:1_Subsample:1	StationVisit:1_Sample:1_Subsample:1_Taxon:106422_1	Width (expressed as pixels) of biological entity specified elsewhere	http://vocab.nerc.ac.uk/collection/P01/current/WDPIXEL1/	10	Pixels	http://vocab.nerc.ac.uk/collection/S02/current/S029/
StationVisit:1_Sample:1_Subsample:1	StationVisit:1_Sample:1_Subsample:1_Taxon:106422_1	Equivalent spherical diameter of biological entity specified elsewhere	http://vocab.nerc.ac.uk/collection/P01/current/OBSINESD/	305	Micrometres (microns)	http://vocab.nerc.ac.uk/collection/P06/current/UMIC/
StationVisit:1_Sample:1_Subsample:2	StationVisit:1_Sample:1_Subsample:1_Taxon:106422_2	Length (expressed as pixels) of biological entity specified elsewhere	http://vocab.nerc.ac.uk/collection/P01/current/LGPIXEL1/	8	Pixels	http://vocab.nerc.ac.uk/collection/P06/current/PIXY/
StationVisit:1_Sample:1_Subsample:2	StationVisit:1_Sample:1_Subsample:1_Taxon:106422_2	Width (expressed as pixels) of biological entity specified elsewhere	http://vocab.nerc.ac.uk/collection/P01/current/WDPIXEL1/	12	Pixels	http://vocab.nerc.ac.uk/collection/S02/current/S029/
StationVisit:1_Sample:1_Subsample:2	StationVisit:1_Sample:1_Subsample:1_Taxon:106422_2	Equivalent spherical diameter of biological entity specified elsewhere	http://vocab.nerc.ac.uk/collection/P01/current/OBSINESD/	329	Micrometres (microns)	http://vocab.nerc.ac.uk/collection/P06/current/UMIC/

For *in-situ* instruments, the fields required are slightly different from benchtop instruments. For example, deployed instruments such as the VPR, do not require the collection of physical samples and calculate the imaged volume using specific formulas with parameters that vary on the settings of the instrument. For example, in Ollevier et. al. (2022) the protocol to calculate plankton densities requires a series of calculations that are specific to the settings used. This information can be stored using *"Sampling protocol"* to specify how the volume was calculated, as shown in Table 4. This information is extracted from the protocol that follows the table.

Table 4: A practical example of sampling acquisition information in the eMoF table for an *in-situ* instrument (VPR).

eventID	*measurementType*	*measurementTypeID*	*measurementValue*	*measurementValueID*	*measurementUnit*	*measurementUnitID*
StationVisit:1	Sampling platform name	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100001/	Simon Stevin	http://vocab.nerc.ac.uk/collection/C17/current/11SS/	not applicable	https://vocab.nerc.ac.uk/collection/P06/current/XXXX/
StationVisit:1	Imaging instrument name	BODC term requested	Video Plankton Recorder {VPR} imaging system - Davis et al. (1992)	http://vocab.nerc.ac.uk/collection/L22/current/TOOL1584/	not applicable	https://vocab.nerc.ac.uk/collection/P06/current/XXXX/
StationVisit:1	Sampling protocol	http://vocab.nerc.ac.uk/collection/P01/current/SAMPPROT/	Imaged volume of a VPR frame= 17.821 ml, computed as the field of view (magnification setting S1: 20.8x15.2 mm) multiplied by focal depth (determined by the parameters used with the VPR AutoDeck software). Sampled volume [ml] = (17.821 [ml/frame] *25 [frames/s] * 10 [s] (Duration of VPR deployment))		not applicable	https://vocab.nerc.ac.uk/collection/P06/current/XXXX/
StationVisit:1_Sample:1	Sample volume	http://vocab.nerc.ac.uk/collection/P01/current/VOLXXXXX/	4455.25		Millilitres	http://vocab.nerc.ac.uk/collection/P06/current/VVML/

Protocol: "During deployment of the VPR, the scientist has to select the VPR's magnification setting and the parameters in the AutoDeck software. The VPR has four preset motor positions that determine the field of view: 8.8x6.6 mm, 20.8x15.2 mm, 33.8x25.5 mm, 46.5x34.5 mm (Seascan, Inc., 2014). These correspond to magnification settings S0 till S3 which are the most zoomed in and zoomed out settings, respectively. The user-defined parameters in AutoDeck are *segmentation threshold – low, segmentation threshold – high, focus – sobel, focus – std dev, growth scale (%), minimum blob size (area) and minimum join distance. The first four of these user-defined parameters in AutoDeck in combination with the magnification setting determine the imaged volume per frame and are calculated by the CalDeck software. With the imaged volume, one can calculate how much water was sampled by the VPR during a transect as is represented in formula 1. To know the sampled volume one has to multiply the imaged volume with the number of fps (for the Real Time VPR this is 25 fps) and the duration that the VPR collects data. In this study, densities are based on the entire trajectory, but densities can also be calculated for a specific part of a trajectory. A shorter deployment time with the respective number of plankton observed within that part of the trajectory should then be used in formula 1 and 2.*

*Sampled volume [ml] = (Imaged volume [ml/frame] *25 [frames/s] * Duration of VPR deployment [s])) (1)*

After validation of the ROIs, the plankton density [ind/m^3] per taxa of a VPR transect can be determined as:

*Density [ind/m^3] = (Number of individuals [ind] / (Sampled volume [ml]) *1,000,000 (2)*

In formula 2 there is a multiplication with 1,000,000 to convert the unit ind/ml to ind/m^3."

4. Workflows: From instrument to EMODnet Biology

In this section we describe the workflow (Figure 4) from imaging acquisition to publication in Biodiversity data platforms in several steps:

1. Raw images and their metadata are retrieved from the instrument,
2. Images are processed with a software (cropping and classification),
    a. This can be done in EcoTaxa (Section 5.1),
3. Derived data is formatted in OBIS-ENV-DATA format,
    a. This format can be exported from EcoTaxa (Section 5.1),
4. Data is submitted to EurOBIS via IPT (Section 5.2),
    a. Quality control by BioCheck tool (Section 6),
5. Data in EurOBIS can flow to EMODnet Biology, OBIS and GBIF.

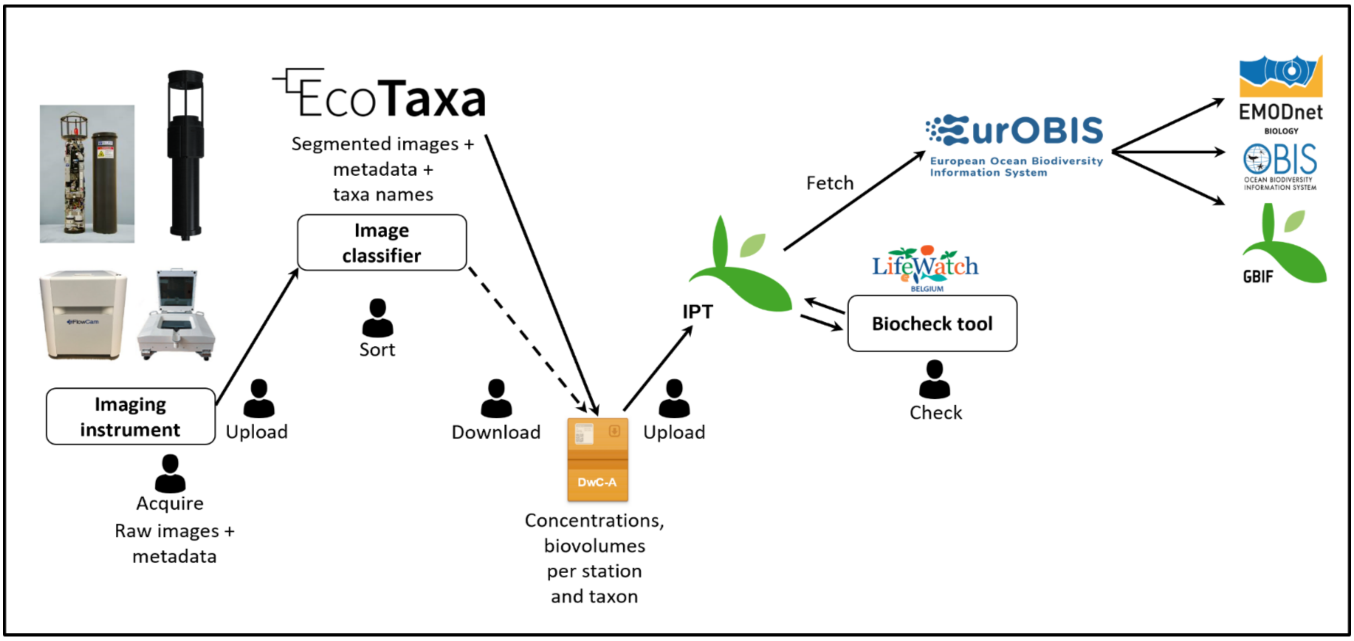
EcoTaxa
Segmented images + metadata + taxa names
Image classifier
Sort
Imaging instrument
Upload
Acquire
Raw images + metadata
Download
DwC-A
Concentrations, biovolumes per station and taxon
Upload
IPT
Fetch
EurOBIS
European Ocean Biodiversity Information System
LifeWatch BELGIUM
Biocheck tool
Check
EMODnet BIOLOGY
OBIS OCEAN BIODIVERSITY INFORMATION SYSTEM
GBIF

Figure 4: Complete workflow from instrument to EMODnet Biology, passing (or by-passing) EcoTaxa to classify the images.

4.1. EcoTaxa workflow

EcoTaxa (http://ecotaxa.obs-vlfr.fr) is a web application that allows users to taxonomically classify images of individual organisms. First, a user needs to upload the data in the application. The import format is a folder with images and a .tsv table with one line per image and many data fields for each (organised hierarchically, from sample to image; which maps well to DwC-A). For many imaging instruments, the processing software can produce a simple format directly . The images and data are stored in a database, within a "project"; a project is a data management unit containing data from a single instrument and over which permissions can be defined for various users. A machine learning model trained on a combination of image features that the user may have uploaded as data with images with features extracted by Convolutional Neural Networks is used to predict a likely identification for each image. The user can review the automatic identifications, validate them in large batches or correct them when needed. Sorting images according to the classifier's confidence score is instrumental in speeding up the review process as well as making it more accurate. Finally, the user can export the data in the same format it was imported in, but with the identifications added.

Furthermore, EcoTaxa can directly export files in the DwC-A format. In that case:

- Users have to create a "collection" of one or several projects. A collection allows to regroup various years of a time series or various legs of a cruise for example, if each year/leg was a separate project. Dedicated DwC-A metadata is defined at collection level (citation, summary, etc.).
- The formula to compute concentrations has to be defined and to be homogeneous within the projects.
- The exported data is aggregated at occurrence level (one taxon in one sample) and the concentration is reported in the EMOF table.

As of today, several of these functionalities are only present in the API and do not yet have a user interface; the development of which is planned in the coming months. Later on, the export to DwC-A of the individual measurements (e.g. ESD) will be considered, once the full data flow chain is able to handle such large datasets.

4.2. EurOBIS harvest and EMODnet Biology publication

The first step to submit the data is to set up an IPT (Integrated Publishing Toolkit) instance in the EurOBIS IPT. IPT is an open source software tool to publish and share datasets, developed by GBIF (Robertson et al., 2014), and adopted by (Eur)OBIS and EMODnet Biology. The IPT software allows users to map their data to the Darwin Core terms and to archive and compress the files as a DwC-A zip file that contains:

- Three data files (in .csv or .txt) related to the three tables (Event, Occurrence and eMoF) of the OBIS-ENV-DATA format,
- a file (eml.xml) containing the metadata of the dataset (see section 8.1),
- a descriptor file (meta.xml) with the different terms used and the relationships between the data files.

The IPT instance can be set at the data provider’s own server. Documentation on how to do this can be found at https://github.com/gbif/ipt/wiki/IPT2ManualNotes.wiki#install-the-ipt. Alternatively, EurOBIS can create an IPT at their own server, this can be requested at info@eurobis.org. After an IPT instance has been set up, the IPT resource should be created to map your data and metadata in IPT. For this, it is recommended to store your data files as tab delimited .txt. This tutorial from GBIF, shows how to add or map the data and metadata in IPT. Once data and metadata have been filled in, the IPT resource needs to be published and consequently the DwC-A dataset is generated. After publication, it is ready to go through a series of quality control (QC) procedures (see section 7).

When the dataset fully complies with the EurOBIS and EMODnet Biology standards, it is ready to be harvested. Harvests in EurOBIS occur every three months in a semi-automated process. The EurOBIS data management team approves and processes the dataset, becoming available in the EurOBIS database and through the viewer and the download toolbox in the EMODnet Biology Portal. From the EurOBIS IPT, OBIS harvests datasets that will be available at http://obis.org/.

5. Quality control

The initial quality control (QC) procedures of the data during data acquisition and the taxonomic identifications are under the responsibility of the person(s) collecting and analysing the data. The QC procedures or protocols can be described in the metadata of the dataset (see section 7.1). In this document, only the QC procedures performed in a DwC-A file, or a dataset published on an IPT, are mentioned and focused on how the dataset meets the EMODnet Biology data quality criteria.

The online tool, LifeWatch-EMODnet Biology QC tool, performs a detailed QC on a OBIS-ENV-DATA dataset, allowing for a visual exploration of the dataset and highlighting potential issues running integrity, format and visual checks. These checks look at, for example if all required fields are present or if the required standards and formats are correct. The tool is based on the EMODnetBioCheck R package and it is available from the LifeWatch services at http://rshiny.lifewatch.be/BioCheck/, where more information on the checks and how to use the tool are explained. Checks are also run by EMODnet Biology during submission. Additionally, there are a number of useful tools developed by OBIS l to correct mistakes and to verify the quality of the datasets. This can be found at https://github.com/iobis/obistools.

With regard to imaging datasets, the new DwC terms recommended to be used in the Occurrence table (section 5.2.2), are planned to be implemented in the QC tool, to be able to assess the quality of dataset based on the field *IdentificationVerificationStatus*, which provides information on the uncertainty of the identification of the organisms reported.

6. Access & use

6.1. Dataset description

The metadata of a dataset consists of the structured information describing the dataset. t. Describing the dataset will help other users to better understand the content and facilitate data discovery and reuse. For this, it is recommended to describe it in a metadata catalogue. Following the workflow presented in this document,, we specify the procedure to submit the metadata of the dataset via EurOBIS, which uses the Ecological Metadata Language (EML) as its metadata standard. Detailed information on the metadata required when submitting a dataset to EurOBIS is found in their online dataset submit form. When a dataset is submitted and accepted in EurOBIS, the metadata is also described and made publicly available in the Integrated Marine Information System (IMIS), developed and hosted by VLIZ. When the data have been made available, the metadata record includes a link to where the data is and where it can be downloaded. This is an example for a dataset described in IMIS: https://www.vliz.be/en/imis?module=dataset&dasid=4687

For imaging datasets we recommend that the metadata includes:

- When possible, add the following in the field “external links” of the metadata:
  - If using EcoTaxa, a URL link to the project where images are stored
  - A URL link to a document with specific details about the report calibration of the instruments, image processing software documentation and any computations (including equations or codes).
  - A URL link to a document with a description of processing methods for automated classifications, including the version of software
- The field “keywords” of the metadata should contain words indicating that the origin of the dataset is imaging. For example: imaging, images, the name of the imaging sampling instrument, etc.

6.2. Dataset distribution and web services

EurOBIS datasets flow to EMODnet Biology, and these are described in the EMODnet Biology Catalogue (e.g. https://www.emodnet-biology.eu/data-catalog?module=dataset&dasid=6505), accessed and viewed via:

- Download toolbox: https://www.emodnet-biology.eu/toolbox/en/download/occurrence/explore
- Web mapper: https://www.emodnet-biology.eu/portal/index.php
- EMODnet Biology API: https://www.emodnet-biology.eu/emodnet-biology-api
- IPT: http://ipt.vliz.be/eurobis/

When datasets have been also shared with OBIS, it is possible to search and download them via:

- OBIS Mapper: https://mapper.obis.org/
- OBIS Web services: https://api.obis.org/
- robis R package: Access using https://obis.org/manual/accessr/

Additionally, if the dataset was also published in GBIF, it can be searched in https://www.gbif.org/.

6.3. Dataset citations

The use of Digital Object Identifiers (DOI) for published datasets is highly recommended. A DOI is a static, permanent link to the dataset, that provides evidence for data claims and allows the dataset to be citable, traceable, and more visible. A DOI can be assigned to a dataset that has been submitted and ingested to EurOBIS. The DOI can be created either by the data provider or by the EMODnet Biology. If the DOI is created by the data provider, it should be done after the QC procedures have been carried out by the EMODnet Biology team. In this case, the "Publisher of the dataset's citation" can be the data provider's organisation or the institution/system that archives the DOI. If the DOI is created by EMODnet Biology, it will be generated when the metadata of the dataset is created in IMIS. In this case, the "Publisher of the dataset's citation" will be the Marine Data Archive (MDA) as that is the publishing instance that will be responsible for the long-term storage of the dataset. If the DOI has been created by EMODnet Biology, and there is an update of the dataset, after creating the new version, the EurOBIS data management team should be informed, and they will run the QC procedure on the new version. This will be re-harvested and a new DOI would be assigned to this new version of the dataset. All the versions of the dataset will be linked in IMIS, the metadata system behind the EMODnet Biology Catalogue.

6.4. Dataset licence

Following FAIR principles, data that flows to the European data portals, is open access and unrestricted. This means that data will be freely accessible at no charge to third parties and available for the long term, as long as the data repository exists. In EMODnet Biology and when filling the metadata in IPT, it is only allowed to choose from three creative commons licences: CC-0 Public Domain Dedication, CC-BY Attribution and CC-BY-NC non-commercial.

7. Data Management Issues

One of the most important issues is the lack of a specialised image repository for long-term archiving of images and image metadata catalogues. As of today, no image repositories dedicated to the storage of images and metadata files specific to plankton exist. The main constraint for the development of such a repository is storage capacity. Imaging data can be highly voluminous, especially in the case of instruments that record data at very fine temporal and spatial resolutions. The retrieval of images from such a repository would also present a problem due to the high data volumes. Consequently, and due to the lack of specialised platforms, the connection between images and their data can be lost during the process from raw data (instrument) to final data repository, or in the worst scenario whole datasets can be omitted from FAIR practices by being stored exclusively locally. There are some facilities capable of storing imaging data and images, for example EcoTaxa. In EcoTaxa, images, data and general information on the sampling process and projects from which the data are connected to, can be stored together. However, the current goal of EcoTaxa is not to become an image repository, but to be a collaborative tool for the annotation of images. We highlight the need for a system (or several systems that are efficiently linked) that facilitates the permanent hosting and storage of images and data, including links to the output content of the data (e.g. links to a dataset and/or data product in a public repository). Additionally, associated provenance documentation should be maximised, including Standard Operating Procedures (SOPs) or calibration reports from the instruments. The repository system should be accessible by both humans and machines, through the inclusion of key words, a graphic user interface, and semantic annotations that provide context for the images at a glance. Imaging instruments create very large datasets very quickly and the capacity to hold these needs to be taken into consideration in the long term.

Other data management issues concern the lack of standard practices or consensus to populate DwC fields such as *identificationVerificationStatus*. This field specifies the status of the reported identification, allowing users to distinguish subsets of data that have been verified by a human (high degree of certainty in the identification of the organism) from those that have not been verified by a human (generally lower degree of certainty). The Darwin Core reference guide recommends the use of controlled vocabularies from HISPID (Herbarium Information Standards and Protocols for Interchange of Data) or ABCD (Access to Biological Collection Data). However, after looking into these in detail, the proposed categories are specific for herbarium specimens, and therefore not fitting the use of these for our case. There are ongoing discussions about this issue in the imaging community, and due to the lack of a standard practice, we feel the need to use non controlled vocabularies, proposing to follow the same terminology currently used in EcoTaxa. Moreover, we propose the use of *identificationReferences*, to capture information about the software or algorithm that aided the identification. Unfortunately the current DwC definition for this field, being "A list (concatenated and separated) of references (publication, global unique identifier, URI) used in the Identification", does not imply technological references. Therefore, a revision to update this definition, considering the commonplace nature of automated identification of organisms through artificial intelligence approaches, would be beneficial. As well as, to include a technological reference to the examples in the DwC term definition.

In our recommendations, we propose optimised ways to fit aggregated data in the current formats required by data repositories. However, a standard practice on how to aggregate the data at the data provider level, before submitting to platforms such as EcoTaxa or directly to the data portals, needs to be established. Some instruments produce very large amounts of near real time data at a high temporal resolution. The IFCB for example can generate up to 30.000 images per hour. Thus, a standard practice based on the (ecological) representativeness of these data among different users of the same instrument needs to be defined.

Another issue deals with reporting non-biological particles into the data portals. However, the scope of EurOBIS is to report only living organisms and their data formats are designed to report taxonomic information. Currently there are some datasets in EurOBIS that contain non-biological data, such as plastics or detritus. However, this information can only be reported in the eMoF table (making use of BODC terms), and not in the occurrence table. In Neeley et al. (2021) a workflow is described to report these data in an external document file that contains non-biological particles standardised names (e.g. detritus, faecal pellet). We highlight the importance of reporting non-biological data because plankton imaging datasets may contain a considerable amount of particulate matter, such as marine snow particles containing organic and inorganic detritus, faecal pellets, and parts of dead organisms. These images can be highly informative in plankton surveys, as the abundance and types of faecal pellets as well as the presence of dead organisms and constructs of planktonic origin, such as mucilaginous larvacean houses, give clues about the plankton community complementary to the images of the organisms themselves (e.g. Robinson et al., 2005; Wilson et al., 2013). Additionally, marine snow imaging data can provide valuable information on oceanic carbon fluxes, as the accumulation of matter into flocs accelerates vertical export rates and modifies planktonic and microbial interactions with sinking organic matter (see e.g. Shanks 2002).

An additional important aspect is that, for this work, several DwC terms that were not required in the OBIS-ENV-DATA format have been added in the EurOBIS database to facilitate ingesting of enriched imaging data. Increasing advancements in collecting plankton observations create the need to advance the way we are managing these data. Most recently, EurOBIS has adapted its harvesting procedures to be able to cope with the proposed data format, and new or updated datasets from EcoTaxa are scheduled to be submitted in the coming months.

To conclude, in the long-term we suggest that (1) the EurOBIS technical infrastructure will continue to be adapted to this format, and foresee that, if widely adopted by the community, large amounts of data may be flowing in their direction in the near future, (2) this format is widely used by the community and adopted by other OBIS nodes, (3) the proposed dataflows using EcoTaxa can be adopted by the organisations generating and/or managing these data, and that EcoTaxa keeps on expanding its capabilities, (4) standard practices on the spatio-temporal aggregation of data are discussed and established, (5) outreach activities to disseminate this information are shared within the wider imaging community, (6) synergies between different projects and working groups in imaging are established (e.g. ITAPINA, Belmont Forum) in order to align the different imaging initiatives and to avoid duplication of work, and (7) stretch collaborations among data generators, data managers from national and international data platforms and working groups of standard bodies are established.

8. Summary

The recommendations and best practices presented here, rather than proposing a new format, make use of the existing Extended Measurement or Facts (eMoF) DwC-A extension from OBIS (De Pooter et. al. 2017). Including additional DwC fields and new BODC vocabularies specific for imaging data, allows to provide important provenance information, improving the interoperability and reusability of these datasets. As technologies evolve, the way and efficiency on how data is being collected is changing, and as such, databases also need to adapt to these needs.

The recommended format for imaging datasets, provides a practical compromise between reporting occurrences at a fine taxonomic level versus abundances at a higher taxon. We therefore propose a method to report quantitative information, along with occurrences data into specialised biodiversity databases, such as EurOBIS. Additionally, and due to the infinite amount of data generated by these instruments and the way the species observed are identified, we also propose a way to report both, the fully automatic but less accurate data and the accurate manual classification of plankton species. These data can be easily subset by users who decide which data better fits their needs.

This format is specifically designed for submission to the European data platforms EurOBIS and EMODnet Biology. However, we envision that it can also be beneficial for submission to OBIS, which also follows the OBIS-ENV-DATA format, and we encourage that the new DwC terms proposed here and recommendations on what data to include in the eMoF table can also be adopted in OBIS.

Finally, we highlight the importance of sharing and making this valuable quantitative imaging data publicly available. Following these recommendations, submissions of datasets originating from different imaging instruments to the European portals can be combined, thus encouraging cross collaborations to create data products covering broader geographic scales and plankton species.

9. Acknowledgments

We acknowledge the contributions from the JERICO, the global marine plankton and the OBIS communities for their support and feedback provided during the course of this work. We thank the British Oceanographic Data Centre (BODC) for the creation of the vocabulary terms necessary for the proposed recommendations.

We would also like to thank Leen Vandepitte, Yi Ming Gan, Pieter Provoost, Stace E. Beaulieu, Margaret O'Brien and Ian Salter for their expertise in this field and their comments that greatly improved this document.

This work was possible thanks to the JERICO-S3 project funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 871153.

10. Annexes

10.1. Imaging instruments and their BODC identifier.

Imaging instrument	BODC identifier
CytoBuoy CytoSense flow cytometer	http://vocab.nerc.ac.uk/collection/L22/current/TOOL1209/
IFCB	http://vocab.nerc.ac.uk/collection/L22/current/TOOL1588/
CPICS	http://vocab.nerc.ac.uk/collection/L22/current/TOOL1582/
UVP 5	http://vocab.nerc.ac.uk/collection/L22/current/TOOL1577/
UVP 6	http://vocab.nerc.ac.uk/collection/L22/current/TOOL1578/
VPR	http://vocab.nerc.ac.uk/collection/L22/current/TOOL1584/
LISST-Holo	http://vocab.nerc.ac.uk/collection/L22/current/TOOL1585/
Zoocam	http://vocab.nerc.ac.uk/collection/L22/current/TOOL1587/
Loki	http://vocab.nerc.ac.uk/collection/L22/current/TOOL1586/
FlowCam	http://vocab.nerc.ac.uk/collection/L22/current/TOOL1583/
FastCam	http://vocab.nerc.ac.uk/collection/L22/current/TOOL1580/
Zooscan	http://vocab.nerc.ac.uk/collection/L22/current/TOOL1581/
Planktonscope	http://vocab.nerc.ac.uk/collection/L22/current/TOOL1579/
ISIIS	http://vocab.nerc.ac.uk/collection/L22/current/TOOL1561/

10.2. Typical measurements derived from imaging and their BODC identifier.

Measurements derived from Imaging	BODC identifier
Length (in digital image) of biological entity specified elsewhere	http://vocab.nerc.ac.uk/collection/P01/current/LGPIXEL1/
Width (in digital image) of biological entity specified elsewhere	http://vocab.nerc.ac.uk/collection/P01/current/WDPIXEL1/
Equivalent spherical diameter of biological entity specified elsewhere	http://vocab.nerc.ac.uk/collection/P01/current/OBSINESD/
Height of pixel	http://vocab.nerc.ac.uk/collection/P01/current/HTPIXEL2/
Width of pixel	https://vocab.nerc.ac.uk/collection/P01/current/WDPIXEL2/
Area of pixel	http://vocab.nerc.ac.uk/collection/P01/current/ARPIXEL2/

References

Colas, F., Tardivel, M., Perchoc, J., Lunven, M., Forest, B., Guyader, G., ... & Romagnan, J. B. (2018). The zoocam, a new in-flow imaging system for fast onboard counting, sizing and classification of fish eggs and metazooplankton. *Progress in Oceanography*, *166*, 54-65. https://doi.org/10.1016/j.pocean.2017.10.014

Cowen, R. K., & Guigand, C. M. (2008). *In situ* ichthyoplankton imaging system (ISIIS): system design and preliminary results. *Limnology and Oceanography: Methods*, *6*(2), 126-132. https://doi.org/10.4319/lom.2008.6.126

Davis, C. S., Thwaites, F. T., Gallager, S. M., & Hu, Q. (2005). A three-axis fast-tow digital Video Plankton Recorder for rapid surveys of plankton taxa and hydrography. *Limnology and Oceanography: Methods*, *3*(2), 59-74. https://doi.org/10.4319/lom.2005.3.59

De Pooter, D., Appeltans, W., Bailly, N., Bristol, S., Deneudt, K., Eliezer, M., ... & Hernandez, F. (2017). Toward a new data standard for combined marine biological and environmental datasets-expanding OBIS beyond species occurrences. *Biodiversity Data Journal*, (5). DOI:10.3897/BDJ.5.e10989

Gorsky, G., Ohman, M. D., Picheral, M., Gasparini, S., Stemmann, L., Romagnan, J.-B., et al. (2010). Digital zooplankton image analysis using the zooscan integrated system. *J. Plankton Res.* 32, 285–303. https://doi.org/10.1093/plankt/fbp124

Lombard, F., Boss, E., Waite, A. M., Vogt, M., Uitz, J., Stemmann, L., ... & Appeltans, W. (2019). Globally consistent quantitative observations of planktonic ecosystems. Frontiers in Marine Science, 196. https://doi.org/10.3389/fmars.2019.00196

MarineRegions.org: Flanders Marine Institute (2022). Available online at www.marineregions.org. Consulted on 2022-03-21.

Neeley, A., Beaulieu, S., Proctor, C., Cetinić, I., Futrelle, J., Soto Ramos, I., Sosik, H., Devred, E., Karp-Boss, L., Picheral, M., Poulton, N., Roesler, C., and Shepherd, A. 2021. Standards and practices for reporting plankton and other particle observations from images. Technical Manual. Woods Hole, MA, Ocean Carbon & Biogeochemistry Project, 38pp. DOI: 10.1575/1912/27377. http://dx.doi.org/10.25607/OBP-1634

Ollevier, A., Mortelmans, J., Vandegehuchte, M.B., De Troch, M., & Deneudt, K. (2022). A Video Plankton Recorder user guide: lessons learned from in-situ plankton imaging in shallow and turbid coastal waters in the Belgian part of the North Sea. (Submitted)

Picheral, M., Guidi, L., Stemmann, L., Karl, D. M., Iddaoud, G., & Gorsky, G. (2010). The Underwater Vision Profiler 5: An advanced instrument for high spatial resolution studies of particle size spectra and zooplankton. *Limnology and Oceanography: Methods*, *8*(9), 462-473. https://doi.org/10.4319/lom.2010.8.462

Picheral, M., Catalano, C., Brousseau, D., Claustre, H., Coppola, L., Leymarie, E., Coindat, J., Dias, F., Fevre, S., Guidi, L., Irissonm J.O., Lombard, F., Mortier, L., Penkerch, C., Rogge, A., Schmechtig, C., Thibault, S., Tixier, T., Waite, A., Stemmann, L. (2021). The Underwater Vision Profiler 6: an imaging sensor of particle size spectra and plankton, for autonomous and cabled platforms. *Limnology and Oceanography: Methods*. 20(2), 115-129. https://doi.org/10.1002/lom3.10475

Pollina, T., Larson, A. G., Lombard, F., Li, H., Colin, S., de Vargas, C., & Prakash, M. (2020). PlanktonScope: affordable modular imaging platform for citizen oceanography. *BioRxiv*. https://doi.org/10.1101/2020.04.23.056978

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779-788.

Robertson, T., Döring, M., Guralnick, R., Bloom, D., Wieczorek, J., Braak, K., ... & Desmet, P. (2014). The GBIF integrated publishing toolkit: facilitating the efficient publishing of biodiversity data on the internet. *PloS one*, *9*(8), e102623. https://doi.org/10.1371/journal.pone.0102623

Robison, B.H., Reisenbichler, K.R., & Sherlock, R.E. (2005). Giant larvacean houses: Rapid carbon transport to the deep sea floor. Science, 308(5728), 1609-1611.DOI: 10.1126/science.1109104

Schulz, J., Barz, K., Ayon, P., Luedtke, A., Zielinski, O., Mengedoht, D., & Hirche, H. J. (2010). Imaging of plankton specimens with the lightframe on-sight keyspecies investigation (LOKI) system. *Journal of the European optical society-rapid publications*, *5*. DOI:10.2971/jeos.2010.10017s

Shanks, A. (2002). The abundance, vertical flux, and still-water and apparent sinking rates of marine snow in a shallow coastal water column. Continental Shelf Research, 22(14), 2045-2064. https://doi.org/10.1016/S0278-4343(02)00015-8

Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, et al., (2012) Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLoS ONE 7(1): e29715. https://doi.org/10.1371/journal.pone.0029715

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, *3*(1), 1-9. https://doi.org/10.1038/sdata.2016.18

Wilson, S. E., Ruhl, H. A., & Smith, Jr, K. L. (2013). Zooplankton fecal pellet flux in the abyssal northeast Pacific: A 15 year time-series study. *Limnology and oceanography*, *58*(3), 881-892. https://doi.org/10.4319/lo.2013.58.3.0881

WoRMS Editorial Board (2022). World Register of Marine Species. Available from https://www.marinespecies.org at VLIZ. Accessed 2022-03-21. https://doi.org/10.14284/170