Ingesting, validating, long-term storage and access of Flow Cytometry data

WP9 - Deliverable D9.13
<table>
<thead>
<tr>
<th>Deliverable number</th>
<th>Short title</th>
</tr>
</thead>
<tbody>
<tr>
<td>D9.13</td>
<td>SDN data management protocols for Flow Cytometer data</td>
</tr>
</tbody>
</table>

**Long title**

Ingesting, validating, long-term storage and access of Flow Cytometer data

**Short description**

This document highlights the way to integrate the flow cytometry datasets into a database that fits interoperability and meets SDC ingesting procedure in order to make the datasets available through international portals. The works presents the first steps of standardisation of vocabulary dedicated to flow cytometry variables (optical units, abundance) and resolved functional groups (phytoplankton and heterotrophic prokaryotes) thanks to a consortium of international experts. The database and workflow from the sample to the SDC portal, ODV and EMODNET are presented, as well as the first datasets available online.

**Author**

CNRS, NERC-BODC, VLIZ and ICES

**Working group**

WP9.5.2

**Dissemination**

Public

**Copyright terms**

Public

### History

<table>
<thead>
<tr>
<th>Version</th>
<th>Authors</th>
<th>Date</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Soumaya Lahbib</td>
<td>12/07/2017</td>
<td>First draft</td>
</tr>
<tr>
<td>2</td>
<td>Soumaya Lahbib</td>
<td>23/05/2018</td>
<td>Second draft</td>
</tr>
<tr>
<td>3</td>
<td>Simon Claus, Paula Oset</td>
<td>04/05/2019</td>
<td>Third draft</td>
</tr>
<tr>
<td>4</td>
<td>Paula Oset</td>
<td>24/06/2019</td>
<td>Fourth draft</td>
</tr>
<tr>
<td>5</td>
<td>Michèle Fichaut</td>
<td>04/04/2019</td>
<td>Final edits</td>
</tr>
<tr>
<td>6</td>
<td>Mélilotus Thyssen</td>
<td>05/07/2019</td>
<td>Final edits</td>
</tr>
</tbody>
</table>
Disclaimer
The content of this document reflects only the authors’ view; it cannot be considered to reflect the view of the European Commission or any other body of the European Union. The European Commission is not responsible for any use that may be made of the information it contains.

Acknowledgment
This document is a deliverable of the SeaDataCloud project. This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement Nº 730960.
Table of contents

1. Background ........................................................................................................................................... 5
2. Methodology ........................................................................................................................................ 5
  2.1. Flow Cytometry common vocabulary establishment ........................................................................ 5
    2.1.1. Critical Analysis of the existing NERC-BODC FCM codes ......................................................... 6
    2.1.2. Captured parameters for Automated FCM .................................................................................. 7
    2.1.3. Literature analysis from 1983 to 2017 ......................................................................................... 8
    2.1.4. Flow cytometry vocabulary standardization questionnaire ....................................................... 9
    2.1.5. FCM Common Vocabulary setting .............................................................................................. 11
  2.1.2. Ingestion, validation and long term access of FCM data .............................................................. 12
  2.1.3. Further integration of FCM data into EMODnet Biology infrastructure ...................................... 13
3. Conclusion ............................................................................................................................................. 16
4. List of acronyms ..................................................................................................................................... 16
5. ANNEX 1- References of the literature review .................................................................................... 18
6. ANNEX 2- Flow Cytometry vocabulary standardisation ....................................................................... 29
  Questionnaire ....................................................................................................................................... 29
7. ANNEX 3- Questionnaire answers ....................................................................................................... 34
8. ANNEX 4- Common vocabulary P02, P01 and F02 lists ........................................................................ 37
1. Background

Marine microbial communities play a major role in the functioning of the global ecosystem. They are good indicators of marine health due to their sensitivity to their environment, and they play a key role in the biogeochemical cycles. Flow cytometry (FCM) is a powerful technology to investigate them. FCM measures the optical properties of single particles (cells) aligned and separated in a laminar flow stream as they cross a light source (most often, one or several laser beams). FCM enables to record various fluorescences intensities produced by the cells, the light scatter intensities per cell, and to determine the abundances of the various groups evidenced. Typically, groups of pico-, nano- and microphytoplankton, heterotrophic prokaryotes, viruses, heterotrophic nanoflagellates are defined by their inherent optical properties. Some specific and very recent instruments are also able to produce real-time, high-resolution data with pictures of each single cell as it flows, giving additional taxonomical identification of cells when larger than about 15 µm.

Flow cytometry manufacturers have harnessed the power of technology to improve their products and offer a wide range of instruments, both for conventional benchtop instruments and for automated flow cytometry deployed in the field, such as: FACS Calibur, BD Influx, CytoSense, FlowCytoBot and recently a CytoPro (with a staining module). These sensors are creating a range of new data types and data formats for which no standards or data management guidelines were available.

FCM data are processed using either the softwares provided by the manufacturers to control the cytomter, or different softwares used for data analysis only such as for instance: FlowJo, Summit, WinList, WinMDI, CytoClus, EasyClus, RtoolClus. Even though the efficiency and the conviviality of these softwares for analyzing and getting results from the acquired measurements are optimised, they deliver different output formats, file schemes and no common vocabulary. Thus, access to standardized and interoperable flow cytometry data were still challenging because of barriers in a common standardized vocabulary definition.

Within the SeaDataCloud project (WP9.5.2), Flow cytometry (FCM) data are considered as new data type that have to be ingested, validated to provide a long-term storage and easy access through SeaDataNet infrastructure. The main objectives of this WP consist of setting up an interoperable system to structure and manage FCM data and metadata (from upstream to downstream services) in coherence with international standards. This work is pioneering in both Flow Cytometry and data management fields.

2. Methodology

This work was carried out from February 2017 to May 2018 (15 Months) where teleconferences and meetings have been established between flow cytometry community and partners. A methodology was established for defining a common set of terms that could be used by a worldwide community of Flow cytometry users. Then, the CNRS-MIO has adapted his FCM local data management method to SeaDataNet tools in order to fulfil the work-package 9.5.2 goals.

2.1.1. Flow Cytometry common vocabulary establishment

This part was realized thanks to the existing conjunction/interaction between SeaDataCloud and JERICO Next projects through their respectively work-packages: WP9.5.2 and Task 3.1 on automated platform for the observation of phytoplankton diversity in relation to ecosystem services. In fact, both projects have a common part that deals with Flow Cytometry vocabulary standardization.
Therefore, the work held on the common vocabulary was as follow:

1. Critical Analysis of the existing NERC-BODC FCM codes
2. Captured parameters exercise
3. Literature review from 1983 to 2017
4. Questionnaire

2.1.1.1. Critical Analysis of the existing NERC-BODC FCM codes

Firstly, a general search through the NERC-BODC vocabulary system (https://www.bodc.ac.uk/resources/vocabularies/vocabulary_search/) by using ‘flow cytometry’ key word. The output (fig.1) revealed that there are 7 vocabulary lists that are holding FCM vocabulary.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Collection</th>
<th>Title</th>
<th>Definition</th>
<th>Governance</th>
<th>Lid</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>P01</td>
<td>BODC Parameter Usage Vocabulary</td>
<td>Terms used in the BODC parameter semantic model designed to describe individual measured phenomena. Must be used to mark up sets of data such as a NERCx area or spreadsheet column.</td>
<td>British Oceanographic Data Centre</td>
<td>📌</td>
</tr>
<tr>
<td>7</td>
<td>L02</td>
<td>Standard Device Catalogue</td>
<td>Terms for distinct sampling or measuring devices that may be identified in the real world in terms of manufacture and model number.</td>
<td>SeaDataNet and MarineUK Vocabulary Central Governance Group</td>
<td>📌</td>
</tr>
<tr>
<td>7</td>
<td>L08</td>
<td>PREMADAS Parameter Usage Vocabulary</td>
<td>Terms under the context governance of TERRAIN used to describe measured phenomena within the PREMADAS project.</td>
<td>Galileo European Space Agency</td>
<td>📌</td>
</tr>
<tr>
<td>4</td>
<td>B01</td>
<td>BODC parameter semantic model analytical method entity descriptions</td>
<td>Controlled vocabulary defining the terms that may be used for an analytical method entity (part of the flow stream) in the BODC parameter semantic model.</td>
<td>British Oceanographic Data Centre</td>
<td>📌</td>
</tr>
<tr>
<td>2</td>
<td>G07</td>
<td>BODC centre parameter collection names</td>
<td>Terms used by BODC to describe groups of related parameters brought together to form a sample from the sample scheme. Each term maps to multiple semantic parameter uses.</td>
<td>British Oceanographic Data Centre</td>
<td>📌</td>
</tr>
<tr>
<td>1</td>
<td>L03</td>
<td>SeaDataNet device catalogue</td>
<td>Terms used to classify groups of sensors, instruments, sources of algorithmically computed data (numerical models), or sensors (collectors of water, SPM, sediment, rock, or bio samples).</td>
<td>SeaDataNet</td>
<td>📌</td>
</tr>
<tr>
<td>1</td>
<td>P01</td>
<td>Global Change Master Directory instrument identifiers</td>
<td>Terms used to describe sensors, instruments and other pieces of scientific equipment in the the Global Change Master Directory metadata.</td>
<td>Global Change Master Directory</td>
<td>📌</td>
</tr>
</tbody>
</table>

Figure 1: The existing list before the new FCM common vocabulary establishment

Secondly, a special attention was given to P01 list about the parameter Usage vocabulary that is commonly used in SeaDataNet. After discussion with the BODC and revision, there were 34 parameter codes related to flow cytometry in the P01 vocabulary.

During the Cytobuoy workshop held in Woerden (The Netherlands) from 27 to 30 March 2017, these parameters were reviewed and discussed between the FCM users (from Euro-Mediterranean laboratories) to identify how much these codes could be helpful.

The feedback pointed out that some codes are good but there was a lot of redundancy and definitions were not clear for the FCM users and difficult to understand. Actually, these have been created over the past 30 years to mark-up datasets received at BODC. Most were created to reflect the terminology used at the source but remodelled to fit the BODC semantic model for biological parameter codes. The collection has grown and increased in diversity over the years as flow cytometry spread in marine laboratories and terminology shifted in response to new experimental applications, greater instrument performance and new scientific understanding. As a result many of these codes became ambiguous, poorly defined, or redundant. This situation is a testimony to the timeliness of agreeing on a set of
common vocabularies and on their definition in order to widely share FCM datasets and make them interoperable with one another.

2.1.1.2. Captured parameters for Automated FCM

In order to upgrade these codes at a broad level of agreement between FCM users, we worked closely with some of the JericoNext partners (CNRS/MIO, Rijkswaterstaat (RWS), the Oceanology and Geosciences laboratory (LOG), VLIZ and the Centre for Environment, Fisheries and Aquaculture Science (Cefas)) on a common exercise to identify their FCM data management method and which parameters are captured after the analysis processing. The result below (fig.2) shows common and unique captured parameters for each partner.

![Figure 2: Synthesis of captured parameters per partner](image)

The combination of all these parameters leads to a total of 73 captured parameters (fig.3) (metadata and data). Since we are focusing on parameter usage vocabulary, our choice was limited to the common 12 data variables found in this exercise based on ‘Area’ criteria (i.e., area of the collected signal). FCM scientists decided to add the same variables based on ‘Height’ criteria (i.e., the peak of the collected signal).
2.1.1.3. Literature analysis from 1983 to 2017

A total of 131 scientific papers (Annex 1) were read starting from the beginning of the flow cytometry technique in the 1980’s till 2017. This literature review allowed to have a thorough understanding on the used instruments, analyses protocols and achieved parameters (Fig.4, 5 and 6).

There are two types of flow cytometers: (i) pulse shape with image in flow recording FCM and (ii) height and area recording FCM. Both of them have different particularities in the number of lasers used to excite the cells and or to trigger the signal emitted by the cells when the cells cross the laser beam. Some automated flow cytometers have the ability to deliver high resolution measurement, in real time. They can be deployed in situ, underwater or on a ship or a buoy. While the more conventional bench top instruments, most of them developed for the biomedical field, are only deployed in the laboratory and mostly analyse samples several days, weeks or months after sampling.
Whatever the instrument type, most of the captured parameters achieved by scientists are about the cells group names, their abundance and the statistics (means, coefficient of variation, etc...) about their optical properties related to Forward scatter, Sideward scatter and fluorescence (orange, red, green, yellow, etc.).

![Figure 4: Evolution of the Flow cytometers](image)

**Figure 4: Evolution of the Flow cytometers**

2.1.1.4. Flow cytometry vocabulary standardization questionnaire

In order to update and/or to create new vocabulary codes with a large consensus of FCM users, a questionnaire (58 questions) was created and submitted to 180 FCM users all over the world. It covers four main parts (fig.6 and annex 2):

**Part I**: Group name and definition => this part includes the captured biological group names and definition based on the literature review. User can add additional group and definition.

**Part II**: Flow Cytometer Metadata => in this part, users were asked about the machine(s), its/their characteristics, signal, sheath fluid and quality control.

**Part III**: Sample Metadata => this part was related to the protocol used during the analyses: standard beads, sample, etc..

**Part IV**: Flow Cytometer Data => This part deals with type of particles, staining, clustering, quality control and the captured parameters.
Figure 6: FCM Questionnaire

After 2 months, 38 answers were collected in which 79% were completed and 21% uncompleted but still usable.

Figure 7: distribution of FCM users who answered the questionnaire

Despite the few number of answers, 90% of the FCM user profiles were composed of researchers and engineers with confirmed to expert levels. Therefore, the collected answers (Annex 3) are so valuable that we succeeded to upgrade and define new common vocabulary codes by involving a large FCM community.
2.1.1.5. FCM Common Vocabulary setting

2.1.1.5.1. P02 and P01 lists

The P01 is related to the BODC Parameter Usage Vocabulary list hosted by the BODC and is based on semantic models whether or not the dataset is chemical or physical or biological (fig.8). In our case, FCM parameters give information on biological as well as non-biological groups of particles (i.e.: Standard fluorescent microsphere used as an internal standard for quantitative and qualitative comparisons). The physical model was chosen in order to avoid the Organism name which is linked to WoRMS and this value is not always guaranteed for all the groups i.e.: ‘Standard beads’, Eukaryote Picophytoplankton, etc... 26 parameters usage vocabulary for FCM needs have been created.

<table>
<thead>
<tr>
<th>Chemical model</th>
<th>Biological model</th>
<th>Physical model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measurement Substance</td>
<td>Measurement Organism Name</td>
<td>Measurement Statistical</td>
</tr>
<tr>
<td>Measurement matrix relationship</td>
<td>Measurement Organism Specifics</td>
<td>Physical entity</td>
</tr>
<tr>
<td>Matrix</td>
<td>Measurement matrix relationship</td>
<td>Measurement matrix relationship</td>
</tr>
<tr>
<td>Matrix subcomponent</td>
<td>Measurement matrix subcomponent</td>
<td>Method</td>
</tr>
<tr>
<td>Concentration of carbon (total inorganic) (TCO2) per unit mass of the water body (dissolved plus reactive particulate phase)</td>
<td>Abundance of Bacteria (ITIS: 202421: WoRMS 6) [Subgroup: heterotrophic] per unit volume of the water body by automated flow cytometry</td>
<td>Area mean of Forward light scatter pulse per cluster from the water body by flow cytometry</td>
</tr>
</tbody>
</table>

The P02 is related to the SeaDataNet Parameter Discovery Vocabulary which is on the top of the P01 group of parameters. In the case of FCM vocabulary defined in the P01, they belong to the ‘FCMW’ code known as ‘Flow cytometry parameters in water bodies’ and defined as the parameters derived from flow cytometry data analysis of water samples using in situ or bench-top flow cytometers (see annex 4).

2.1.1.5.2. F02 – SeaDataCloud Flow Cytometry Standardised Cluster Names

The F02 list was created within the SeaDataCloud project in order to manage all the optical cluster names and definitions identified by FCM. Currently, the list contains 11 codes and can be further extended and upgraded (see annex 4).

http://seadatanet.maris2.nl/v_bodc_vocab_v2/search.asp?lib=F02

2.1.1.5.3. L22 – SeaVoX Device Catalogue

The SeaVoX device catalogue list defines and describes all the devices used for sea measurements. For FCM, we have added 2 devices to the existing list such as: BD FACSCalibur Flow Cytometer and CytoSense flow cytometer.
2.1.1.5.4. P06 – BODC data storage units
In this list we have added the international system unit related to Number per cubic centimetre (NCM3). Equivalent to number per millilitre.

2.1.2. Ingestion, validation and long term access of FCM data
Before ingesting the FCM data, a new FCM SeaDataNet Ocean Data View (ODV) data transport format was created. As both physical and biological, FCM data were not suitable for the existing standardized SeaDataNet transport formats. Here is the case of the CNRS-MIO demonstrating the FCM data management method from the instrument acquisition to the SeaDataNet ingestion (fig.9):

For a dedicated project (a cruise for instance), data files acquired by flow cytometry are analysed through a batch process clustering, converted and validated through the CytoBase Input Processor (a standalone software built on R programme by Mathilde Dugenne). Then, data integration into CYTOBASE (local database) is processed automatically using Talend (Extract Transform and Load (ETL) tool). Subsequently, CYTOBASE is connected to MIKADO (SeaDataNet tool for metadata production) to generate the Common Data Index (CDI) dataset (aggregation of measurements), the Cruise Summary Report (CSR), the European Directories of Marine Environmental Datasets (EDMED) and Marine Environmental Research Projects (EDMERP). Also, the connexion between MIKADO and CYTOBASE allows the generation of the coupling table which is the association of the CDIs (metadata) and the physical data files. Finally, The CDI and the coupling table are sent to SeaDataNet support team (cdi-support) for validation and ingestion into SeaDataNet infrastructure. The connexion between the data centre and the SeaDataNet Request Status Management service (RSM) is made thanks to the Download Manager (SeaDataNet java component tool) which was installed by the CNRS-MIO and has been operating since February 27th, 2018.

Figure 9: Flow Cytometry (FCM) data ingestion into SeaDataNet infrastructure (case of CNRS-MIO)
2.1.3. Further integration of FCM data into EMODnet Biology infrastructure

FCM data was integrated in EMODnet Biology by making use of the Darwin Core (DwC) Event schema, recently adopted by the marine biodiversity community (De Pooter et al, 2017). The DwC Event Core was implemented by OBIS through the OBIS-ENV-DATA project to respond to the growing needs to provide environmental data together with the species occurrences, and to enhance interoperability of the data through the adoption of controlled vocabularies in the extended Measurements or Facts (eMoF) extension. This schema provides the necessary flexibility to include any kind of data that can be linked either to a species occurrence (e.g. biomass or development stage), or to a sampling event (e.g. sample size, temperature of the water).

The increased flexibility allows fitting the FCM data into the DwC Event schema, by making use of the controlled vocabularies developed during the SDC project described in this deliverable. Two test datasets collected as part of the JERICO-Next JRAP actions were integrated in EMODnet Biology and are now available via the download toolbox, or in the following links:


Plankton biodiversity data from a North Sea Cruise with R/V Simon Stevin in May 2017: made available directly in EMODnet Biology.

The OBIS-ENV-DATA contains a table for Extended Measurements or Facts (known as “eMoF extension”) where additional data related to a sampling event or an occurrence can be provided by making use of controlled vocabularies. Using this approach, the FCM data can be provided, linking to the Event and Occurrences IDs by using the developed FCM vocabularies as follows:

Standardised cluster names:

<table>
<thead>
<tr>
<th>eventID</th>
<th>occurrenceID</th>
<th>measurementType</th>
<th>measurementTypeID</th>
<th>measurementValue</th>
<th>measurementValueID</th>
</tr>
</thead>
<tbody>
<tr>
<td>LL_SimonStevin_sws 15flr_2uls_360sec_2017-05-08_15h01.cyz</td>
<td>SimonStevin_08/05/2017_111</td>
<td>Registered name identifier (...)</td>
<td><a href="http://vocab.nerc.ac.uk/collection/P01/current/ID">http://vocab.nerc.ac.uk/collection/P01/current/ID</a> CLFL02</td>
<td>Eukaryote picophytoplankton</td>
<td><a href="http://vocab.nerc.ac.uk/collection/F02/current/F0200004/">http://vocab.nerc.ac.uk/collection/F02/current/F0200004/</a></td>
</tr>
<tr>
<td>LL_SimonStevin_sws 15flr_2uls_360sec_2017-05-08_15h01.cyz</td>
<td>SimonStevin_08/05/2017_112</td>
<td>Registered name identifier (...)</td>
<td><a href="http://vocab.nerc.ac.uk/collection/P01/current/ID">http://vocab.nerc.ac.uk/collection/P01/current/ID</a> CLFL02</td>
<td>Eukaryote nanophytoplankton</td>
<td><a href="http://vocab.nerc.ac.uk/collection/F02/current/F0200005/">http://vocab.nerc.ac.uk/collection/F02/current/F0200005/</a></td>
</tr>
<tr>
<td>LL_SimonStevin_sws 15flr_2uls_360sec_2017-05-08_15h01.cyz</td>
<td>SimonStevin_08/05/2017_113</td>
<td>Registered name identifier (...)</td>
<td><a href="http://vocab.nerc.ac.uk/collection/P01/current/ID">http://vocab.nerc.ac.uk/collection/P01/current/ID</a> CLFL02</td>
<td>Microphytoplankton</td>
<td><a href="http://vocab.nerc.ac.uk/collection/F02/current/F0200008/">http://vocab.nerc.ac.uk/collection/F02/current/F0200008/</a></td>
</tr>
</tbody>
</table>

Table 1. Example of records in the eMoF extension for cluster names. MeasurementTypeID and measurementValueID contain the BODC controlled vocabularies (P01 for measurementTypeID and the developed F02 for measurementValueID). measurementType and measurementValue are free text fields but these are completed with the corresponding vocabulary preferred labels.

Optical properties:
<table>
<thead>
<tr>
<th>eventID</th>
<th>occurrenceID</th>
<th>measurement Type</th>
<th>measurementTypeID</th>
<th>measurement ValueID</th>
<th>measurement Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHROME_S1_2016-03-24T12:02:00.000</td>
<td>CHROME_S1_2016-03-24T12:02:00.000_1</td>
<td>Area mean of red fluorescence pulse per cluster from the water body by flow cytometry</td>
<td><a href="http://vocab.nerc.ac.uk/collection/P01/current/FLRAREA">http://vocab.nerc.ac.uk/collection/P01/current/FLRAREA</a></td>
<td>12241.5704</td>
<td></td>
</tr>
<tr>
<td>CHROME_S1_2016-03-24T12:02:00.000</td>
<td>CHROME_S1_2016-03-24T12:02:00.000_1</td>
<td>Area standard deviation of red fluorescence pulse per cluster from the water body by flow cytometry</td>
<td><a href="http://vocab.nerc.ac.uk/collection/P01/current/FLRARESD">http://vocab.nerc.ac.uk/collection/P01/current/FLRARESD</a></td>
<td>6394.2</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Example of records for FCM optical properties in the eMoF extension using the P01 vocabularies developed.

Other measurements:

<table>
<thead>
<tr>
<th>eventID</th>
<th>occurrenceID</th>
<th>measurement Type</th>
<th>measurementTypeID</th>
<th>measurement ValueID</th>
<th>measurement Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHROME_MARS2016_FCMW</td>
<td></td>
<td>Sampling platform name</td>
<td><a href="http://vocab.nerc.ac.uk/collection/Q01/current/Q0100001/">http://vocab.nerc.ac.uk/collection/Q01/current/Q0100001/</a></td>
<td>Carthage</td>
<td></td>
</tr>
<tr>
<td>CHROME_S1_2016-03-24T12:02:00.000</td>
<td>CHROME_S1_2016-03-24T12:02:00.000_1</td>
<td>Abundance of biological entity (...)</td>
<td><a href="http://vocab.nerc.ac.uk/collection/P01/current/SDBIOL01">http://vocab.nerc.ac.uk/collection/P01/current/SDBIOL01</a></td>
<td>87.69</td>
<td></td>
</tr>
<tr>
<td>CHROME_S1_2016-03-24T12:02:00.000</td>
<td></td>
<td>Volume sampled of the water body</td>
<td><a href="http://vocab.nerc.ac.uk/collection/P01/current/VOLWBSMP/">http://vocab.nerc.ac.uk/collection/P01/current/VOLWBSMP/</a></td>
<td>0.376328</td>
<td></td>
</tr>
<tr>
<td>CHROME_S1_2016-03-24T12:02:00.000</td>
<td></td>
<td>Sampling instrument name</td>
<td><a href="http://vocab.nerc.ac.uk/collection/Q01/current/Q0100002/">http://vocab.nerc.ac.uk/collection/Q01/current/Q0100002/</a></td>
<td>CytoBuoy CytoSense flow cytometer</td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Additional standardised measurements in the eMoF extension.

The underlying data system in EMODnet Biology stores the data following the DwC schema and associated standards, making it interoperable with (Eur)OBIS. However, to increase simplicity for users, the data is flattened when accessed via the EMODnet Biology download toolbox.
The main remaining challenge for marine biodiversity data repositories, such as EMODnet Biology and OBIS, remains how to deal with non-taxonomically resolved records in the searching interfaces. EMODnet Biology has recently launched a new version of its download toolbox, which allows searching
for data standardized to controlled vocabularies in the eMoF extension. It is therefore possible to search specifically for FCM data in EMODnet Biology. However, further investigation and consultation are recommended to find the optimal data storing and querying solutions that can meet the global biodiversity and FCM community needs in the long term.

3. Conclusion

FCM technology and mainly the automated flow cytometers are revolutionizing the biological world by acquiring high resolution (in time and space) and real-time data about the first levels of the marine food web. Making these data sustainable, accessible and standardized will be very useful for the marine community as interoperability will greatly facilitate inter-community discussions. There is still a continuous effort to update and/or define common vocabulary, add new metadata and ingest data into SeaDataNet.

This work was made with a strong interaction between FCM users and scientists from Euro-mediterranean institutes. Scientists have showed a big interest on sharing their data and put them accessible within SeaDataNet portal.

Thanks to Cytobuoy workshop (March 2017), JericoNext workshops (WP3 in 2016 and 2018) and Euormarine (March 2018) on improving the visibility of ocean data from new technologies: a case study of high frequency flow cytometry, we could disseminate SeaDataNet activity through this work-package.

4. List of acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>BODC</td>
<td>British Oceanographic Data Centre</td>
</tr>
<tr>
<td>CDI</td>
<td>Common Data Index (SeaDataNet catalogue)</td>
</tr>
<tr>
<td>CEFAS</td>
<td>The Centre for Environment, Fisheries and Aquaculture Science</td>
</tr>
<tr>
<td>CNRS</td>
<td>Centre National de la Recherche Scientifique (France)</td>
</tr>
<tr>
<td>CSR</td>
<td>Cruise Summary Report (SeaDataNet Catalogue)</td>
</tr>
<tr>
<td>EDMED</td>
<td>European Directory of Marine Environmental Data sets (SeaDataNet catalogue)</td>
</tr>
<tr>
<td>EDMERP</td>
<td>European Directory of Marine Environmental Research Projects (SeaDataNet catalogue)</td>
</tr>
<tr>
<td>ETL</td>
<td>Extract Transform and Load</td>
</tr>
<tr>
<td>FACS</td>
<td>Fluorescence-Activated Cell Sorting</td>
</tr>
<tr>
<td>FCM</td>
<td>Flow Cytometry</td>
</tr>
<tr>
<td>JERICO</td>
<td>Joint European Research Infrastructure Network for Coastal Observatories</td>
</tr>
<tr>
<td>JRAP</td>
<td>Joint Research Activity Projects</td>
</tr>
<tr>
<td>LOG</td>
<td>Oceanology and Geosciences laboratory</td>
</tr>
<tr>
<td>MIO</td>
<td>Mediterranean Institute of Oceanography</td>
</tr>
<tr>
<td>NERC</td>
<td>Natural Environment Research Council</td>
</tr>
<tr>
<td>OBIS</td>
<td>Ocean Biogeographic Information System</td>
</tr>
<tr>
<td>ODV</td>
<td>Ocean Data View</td>
</tr>
<tr>
<td>RSM</td>
<td>Request Status Manager (SeaDataNet service)</td>
</tr>
<tr>
<td>RWS</td>
<td>Rijkswaterstaat, Netherlands</td>
</tr>
<tr>
<td>SDC</td>
<td>SeadataCloud</td>
</tr>
<tr>
<td>SDN</td>
<td>SeaDataNet</td>
</tr>
<tr>
<td>VLIZ</td>
<td>Flanders Marine Institute</td>
</tr>
</tbody>
</table>
5. ANNEX 1- References of the literature review


Gerath, M.W., Chisholm, S.W., 1989. Change in Photosynthetic Capacity over the Cell Cycle in Light/Dark-Synchronized Amphidinium carteri Is Due Solely to the Photocycle. PLANT PHYSIOLOGY 91, 999–1005. doi:10.1104/pp.91.3.999


Girault, M., Arakawa, H., Barani, A., Ceccaldi, H.J., Hashihama, F., Kinouchi, S., Gregori, G., 2013. Distribution of ultraphytoplankton in the western part of the North Pacific subtropical gyre
during a strong La Niña condition: relationship with the hydrological conditions. Biogeosciences 10, 5947–5965. doi:10.5194/bg-10-5947-2013


Olson, R.J., Vaulot, D., Chisholm, S.W., 1986. Effects of Environmental Stresses on the Cell Cycle of Two Marine Phytoplankton Species. PLANT PHYSIOLOGY 80, 918–925. doi:10.1104/pp.80.4.918


6. ANNEX 2- Flow Cytometry vocabulary standardisation

Questionnaire

[Image of Flow Cytometry vocabulary standardisation Questionnaire]

This questionnaire is dedicated to set up a common standardized vocabulary of the flow cytometry (FCM) metadata and data. It will take approximately 15 minutes to be completed.

This questionnaire is carried out within the framework of SeaDataCloud H2020 project in order to standardize, validate and guarantee a long-term storage and access of flow cytometry datasets.

Part I: FCM Group names and definitions

Based on literature from 1982 to 2017, do you agree on these FCM group definitions:

- **Prochlorococcus**
  
  Prochlorococcus cells are defined as the small-sized cyanobacteria found in the marine environment. No staining is required to distinguish them by flow cytometry. The Prochlorococcus cluster has higher FWS and red fluorescence (FLR) signatures than Prochlorocyanales and a distinct orange fluorescence (FLO) signature from their phycocyanins accessory pigment when excited by lasers whose wavelength is below 535 nm. Prochlorocyanales may contain phycocyanins, excited by a red laser and constant shear sonication, which is not seen in the Prochlorococcus group. Due to their small size (0.6-1.2 μm) as reported in the literature, cyanobacterial cells exhibit a low intensity of FWS, 568 and FLR signals. They are ubiquitous phototrophic cyanobacteria with a flow-cytometry forward-scatter (FWS) and side-scatter (SSW) signatures that are larger than those of most of the marine heterotrophic bacteria. No staining is required to distinguish them by flow cytometry. The related cluster has higher FWS and red fluorescence (FLR) signatures than Prochlorocyanales and a distinct orange fluorescence (FLO) signature from their phycocyanins accessory pigment when excited by lasers whose wavelength is below 535 nm. Prochlorocyanales may contain phycocyanins, excited by a red laser and constant shear sonication, which is not seen in the Prochlorococcus group. Due to their small size (0.6-1.2 μm) as reported in the literature, cyanobacterial cells exhibit a low intensity of FWS, 568 and FLR signals.

Check any that apply

- I agree
- I disagree
- **Eukaryotes**

  **Phaeophyta**
  The Phaeophyta (brown algae) are defined as unicellular or multicellular algae with brown pigments. They are found in marine and fresh water environments. Phaeophyta are characterized by the presence of chlorophylls and carotenoids, which give them their brown color. They play a crucial role in the marine ecosystem, serving as primary producers.

  **Check any that apply**
  - I agree
  - I do not agree

- **Eukaryote**

  **Nanophyta**
  The Nanophyta (microscopic algae) are defined as unicellular or small multicellular algae that are less than 20 microns in length. They are found in marine and fresh water environments. Nanophyta are characterized by the presence of chlorophylls and carotenoids, which give them their green color. They play a crucial role in the marine ecosystem, serving as primary producers.

  **Check any that apply**
  - I agree
  - I do not agree

- **Cryptophyta**

  Cryptophyta are defined as unicellular or small multicellular algae that are less than 20 microns in length. They are found in marine and fresh water environments. Cryptophyta are characterized by the presence of chlorophylls and carotenoids, which give them their green color. They play a crucial role in the marine ecosystem, serving as primary producers.

  **Check any that apply**
  - I agree
  - I do not agree

- **Coccolithophores**

  Coccolithophores are defined as unicellular or small multicellular algae that are less than 20 microns in length. They are found in marine and fresh water environments. Coccolithophores are characterized by the presence of chlorophylls and carotenoids, which give them their green color. They play a crucial role in the marine ecosystem, serving as primary producers.

  **Check any that apply**
  - I agree
  - I do not agree

- **Microphytobenthos**

  The microphytobenthos (microalgae) are defined as unicellular or small multicellular algae that are less than 20 microns in length. They are found in marine and fresh water environments. Microphytobenthos are characterized by the presence of chlorophylls and carotenoids, which give them their green color. They play a crucial role in the marine ecosystem, serving as primary producers.

  **Check any that apply**
  - I agree
  - I do not agree

- **heterotrophic bacteria**

  Heterotrophic bacteria are defined as unicellular or small multicellular algae that are less than 20 microns in length. They are found in marine and fresh water environments. Heterotrophic bacteria are characterized by the presence of chlorophylls and carotenoids, which give them their green color. They play a crucial role in the marine ecosystem, serving as primary producers.

  **Check any that apply**
  - I agree
  - I do not agree

- **Standard beads**

  Standard beads are defined as unicellular or small multicellular algae that are less than 20 microns in length. They are found in marine and fresh water environments. Standard beads are characterized by the presence of chlorophylls and carotenoids, which give them their green color. They play a crucial role in the marine ecosystem, serving as primary producers.

  **Check any that apply**
  - I agree
  - I do not agree

Please enter other group(s) name and definition if the above list is not complete:

Group 1:
Group 2:
GROUP 3:
Group 4:
### What model of Flow Cytometer(s) do you use?

- [ ] Eppendorf FL200
- [ ] Beckman Coulter
- [ ] Other: 

### Does your instrument have an image in flow device?

- [ ] Yes
- [ ] No

### What is your sample inlet internal diameter (in microns)?

- [ ] Separate multiple entries with commas

### Which lasers wavelengths do you use (in nm)?

- [ ] Separate multiple entries with commas

### Laser beam powers (in mW)?

- [ ] Separate multiple entries with commas

### How many light scatters does your instrument record?

- [ ] Separate multiple entries with commas

### How many channels does your instrument record?

- [ ] Separate multiple entries with commas

### For each laser, please indicate each optical filter configuration (light scatter, fluorescence) in nanometers?

- [ ] Separate multiple entries with commas

### What signal do you use as trigger?

- [ ] Separate multiple entries with commas

### What type of signal does your instrument record?

- [ ] Check one that apply
  - Pulse channel/TOF
  - Area/Height
  - Width/Max
  - Other: 

### What is the type/Composition of the sheath fluid you use?

- [ ] Check one that apply
  - Natural sea water
  - Artificial sea water
  - Distilled water
  - PBS
  - Other: 

### Do you perform quality control of your instrument?

- [ ] Yes
- [ ] No
### Part III: Sample Metadata

#### What biosensor reference do you use?

- [ ] [e.g. (model, via, manufacturer)]

#### What biosensor diameters do you use?

- [ ] [e.g. 1.5, 3.6, 9 mm, ...]

#### What are your biosensor fluorescences?

- [ ] [e.g. Yellow]

#### For what purpose do you use this instrument?

Check any that apply:
- [ ] Monitoring
- [ ] Biotechnology
- [ ] Other:

#### Where is your area of study?

- [ ] [e.g. North sea, North Channel, etc...]

#### What type of sample do you analyze?

Check any that apply:
- [ ] Sea water
- [ ] Fresh water
- [ ] Cultures
- [ ] Other:

#### What is your approximate analyzed volume (ml)?

- [ ] [e.g. 1-10000 ml]

In case of multiple entries, separate with commas.

#### What is your sample flow rate (ml/min)?

- [ ]
Do you use a fluorescent dye?
- Yes
- No

Which type of particles do you measure?
- Phytoplankton
- Mesozooplankton
- Bacteria
- Viruses
- Other: [ ]

What are the recurrent autotrophic functional groups of your area of study?
- Synechococce
- Phaeocystis
- Prochlorococce
- Picoeukaryotes
- Nanoeukaryotes
- Coccolithophore
- Cryptophytes
- Unknown
- Not concerned
- Other: [ ]

What are the recurrent heterotrophic groups of your area of study?
- High Nucleic Acid Prokaryotes
- Low Nucleic Acid Prokaryotes
- Nanoflagellates
- Not concerned
- Other: [ ]

What clustering method do you use?
- Manual
- Automatic
- Other: [ ]

Do you flag your data?
- Yes
- No

What parameters do you export after your clustering?
- Functional group names
- Abundance (cell cm^-3)
- Average Side Ward Scatter (Area, length)
- Average Forward Scatter (Area, length)
- Average Mid Pixeliances
- Average Orange Fluorescence
- Standard deviation Side Ward Scatter (Area, length)
- Standard deviation Forward Scatter (Area, length)
- Standard deviation Had Fluorescence
- Standard deviation Orange Fluorescence
- Other: [ ]

What is the unit used for scatter and fluorescence?
- Arbitrary unit (a.u.)
- Other: [ ]
7. ANNEX 3- Questionnaire answers

**PART I: Groups definition from the FCM point of view**

Q: Based on literature from 1983 to 2017, do you agree on these group definitions:

- Prochlorococcus
- Synechococcus
- Eukaryotes Picophytoplankton
- Eukaryote Nanophytoplankton
- Cryptophytes
- Coccolithophores
- Microphytoplankton
- Heterotrophic Bacteria

![Pie chart showing 87% agreement and 13% disagreement]

**Suggestion to add other group(s) name**

- Phaeocystis
- Viruses (V1, V2, V3)
- Orange fluorescing microplankton (Dinophysis spp.)
- Elongated cells (high red fluorescence with respect to light scattering)
- Algal viruses
- Virus-small
- Heterotrophic nanoflagellates HF
- Nanocyanobacteria - filaments
- Green fluorescing nanoplanckton (certain heterotrophic dinoflagellates)
- Virus - large
- Green fluorescing microplankton (certain heterotrophic dinoflagellates)
- Green fluorescing microplankton (certain heterotrophic dinoflagellates)
PART II: Flow Cytometer Metadata

Q: What model of Flow Cytometer(s) do you use?

Q: What type of signal does your instrument record?

Q: Does your instrument have an image in flow device?

Q: What is the type/composition of the sheath fluid you use?

Q: What signal do you use as trigger?

PART III: Sample Metadata

Q: How do you make the quality control?

Q: What type of sample do you analyze?

Q: For what purpose do you use this instrument?

Q: What beads and diameters reference do you use?
PART III: Sample Metadata

Q: Which parameters do you export after your clustering?

- Abundance (cell.cm^-3)
- Average Red Fluorescences
- Average Orange Fluorescences
- Average Side Ward Scatter (Area, length)
- Average Forward Scatter (Area, length)
- Other

- Functional group names
- Standard deviation Red Fluorescences
- Standard deviation Orange Fluorescences
- Standard deviation Side Ward Scatter (Area, length)
- Standard deviation Forward Scatter (Area, length)
8. ANNEX 4- Common vocabulary P02, P01 and F02 lists

<table>
<thead>
<tr>
<th>ConceptID</th>
<th>Pref Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASAM</td>
<td>Acoustic backscattering in the water column</td>
</tr>
<tr>
<td>GP080</td>
<td>Fishing by-catch</td>
</tr>
<tr>
<td>FEFF</td>
<td>Fishing effort</td>
</tr>
</tbody>
</table>

**FCMW**: Flow cytometry parameters in water bodies

<table>
<thead>
<tr>
<th>ConceptID</th>
<th>Pref Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>FLGAREA</td>
<td>Area mean of green fluorescence pulse per cluster from the water body by flow cytometry</td>
</tr>
<tr>
<td>FLGARES</td>
<td>Area standard deviation of green fluorescence pulse per cluster from the water body by flow cytometry</td>
</tr>
<tr>
<td>FLGMAXA</td>
<td>Peak height mean of green fluorescence pulse per cluster from the water body by flow cytometry</td>
</tr>
<tr>
<td>FLGMAXSD</td>
<td>Peak height standard deviation of green fluorescence pulse per cluster from the water body by flow cytometry</td>
</tr>
<tr>
<td>FLOAREA</td>
<td>Area mean of orange fluorescence pulse per cluster from the water body by flow cytometry</td>
</tr>
<tr>
<td>FLOARES</td>
<td>Area standard deviation of orange fluorescence pulse per cluster from the water body by flow cytometry</td>
</tr>
<tr>
<td>FLOMAXA</td>
<td>Peak height mean of orange fluorescence pulse per cluster from the water body by flow cytometry</td>
</tr>
<tr>
<td>FLOMAXSD</td>
<td>Peak height standard deviation of orange fluorescence pulse per cluster from the water body by flow cytometry</td>
</tr>
<tr>
<td>FLRAREA</td>
<td>Area mean of red fluorescence pulse per cluster from the water body by flow cytometry</td>
</tr>
<tr>
<td>FLRARES</td>
<td>Area standard deviation of red fluorescence pulse per cluster from the water body by flow cytometry</td>
</tr>
<tr>
<td>ConceptID</td>
<td>Preferred label</td>
</tr>
<tr>
<td>-----------</td>
<td>----------------</td>
</tr>
<tr>
<td>IDCLFL02</td>
<td>Registered name identifier of flow cytometry cluster by classification to a term from the NVS SeaDataCloud Flow Cytometry Standardised Cluster Names Vocabulary (SDN-F02:</td>
</tr>
<tr>
<td>NMCLFL02</td>
<td>Registered name of flow cytometry cluster by classification to a term from the NVS SeaDataCloud Flow Cytometry Standardised Cluster Names Vocabulary (SDN:F02:</td>
</tr>
<tr>
<td>ConceptID</td>
<td>Preferred label</td>
</tr>
<tr>
<td>----------</td>
<td>-------------------</td>
</tr>
<tr>
<td>F0200001</td>
<td>Standard beads</td>
</tr>
<tr>
<td>F0200002</td>
<td>Prochlorococcus</td>
</tr>
<tr>
<td>ID</td>
<td>Description</td>
</tr>
<tr>
<td>-------</td>
<td>-------------</td>
</tr>
<tr>
<td>P200003</td>
<td><strong>Synechococcus</strong></td>
</tr>
<tr>
<td>P200004</td>
<td><strong>Eukaryote picophytoplankton</strong></td>
</tr>
<tr>
<td>F0200005</td>
<td>Eukaryote nanophytoplankton</td>
</tr>
<tr>
<td>F0200006</td>
<td>Cryptophytes</td>
</tr>
<tr>
<td>ID</td>
<td>Sample</td>
</tr>
<tr>
<td>-------</td>
<td>-------------------------</td>
</tr>
<tr>
<td>R0200007</td>
<td>Coccolithophores</td>
</tr>
<tr>
<td>R0200008</td>
<td>Microphytoplankton</td>
</tr>
</tbody>
</table>
Heterotrophic prokaryotes include both bacteria and Archaea. They do not contain any photosynthetic pigments and thus do not have any autofluorescence properties exploitable by flow cytometry. Thus, they require a staining with some fluorescent dye to be resolved by flow cytometry. In most studies a nucleic acid dye is used. Staining of nucleic acids by a dye emitting in the green when excited by a blue laser enables heterotrophic prokaryotes to be distinguished in various groups thanks to sideward scatter (or FWS) and FLG signatures:

- Cells with a lower FLG correspond to heterotrophic prokaryotes with a Lower Nucleic Acid content (LNA) and cells with a higher FLG correspond to a Higher Nucleic Acid content (HNA). Their scatter signals (FWS, SWS) are lower than those of Synechococcus and eukaryotic phytoplankton, and may overlap those of Prochlorococcus.

Heterotrophic prokaryote cells with a Higher Nucleic Acid (HNA) content as defined by flow cytometry, based on the intensity of their fluorescence signal induced by a nucleic acid dye.

Heterotrophic prokaryote cells with a Lower Nucleic Acid (LNA) content as defined by flow cytometry, based on the intensity of their fluorescence signal induced by a nucleic acid dye.