Interoperability of new data type with SeaDataNet Infrastructure: Case of Flow cytometry data

Soumaya LAHBIB, MIO (France), soumaya.lahbib@mio.osupytheas.fr
Maurice LIBES, OSUPYTHEAS (France), maurice.libes@osupytheas.fr
Gwenaëlle Moncoiffé, BODC (UK), gmon@bodc.ac.uk
Gerald Gregori, MIO (France), gerald.gregori@mio.osupytheas.fr
Michèle Fichaut, IFREMER (France), michele.fichaut@ifremer.fr
Dick Schaap, MARIS (Netherlands), dick@maris.nl
Mathilde Dugenne, OSU (USA), dugennem@oregonstate.edu
Michel Denis, MIO (France), michel.denis@mio.osupytheas.fr
Pierre Marrec, MIO (France), pierre.marrec@mio.osupytheas.fr
Melilotus Thyssen, MIO (France), melilotus.thyssen@mio.osupytheas.fr

Background and objectives :

Planktonic microbial communities play a major role in the functioning of the global ecosystem. They are good indicators of marine health due to their key role in biogeochemical cycles. Flow cytometry is a powerful technology to investigate them.

Flow Cytometry (FCM) technology measures the optical properties of single (particles) cells in alignment as they cross a light source excitation. It enables to record various fluorescences and light scattering intensities per cell, and determine the abundances of various groups. Typically, groups of pico-, nano- and microphytoplankton, heterotrophic prokaryotes, viruses, heterotrophic nanoflagellates are defined by their inherent optical properties. Some specific instruments are also able to take a picture of the single cell as it flows, giving additional taxonomical identification of cells larger than 20 µm.

This work is part of the SeaDataCloud project (H2020) whose goal is to ingest, validate and provide a long-term storage and easy access to FCM data. Through the SeaDataNet standardized system for marine and ocean data management, FCM datasets will adopt common vocabularies, ISO 19115 metadata standard, Data Transport Formats along with protocols for quality control.

Methods :

Before ingesting these data, a special focus was given to the FCM common vocabulary in order to identify the most used metadata and data. First, a brainstorming work was done within small groups of scientists to come up with the most captured FCM parameters (variables). Then, a bibliography work was carried out so as to find the appropriate definitions and descriptions of these FCM parameters. Finally, a dedicated questionnaire containing 56 questions had for objective to enlarge the impact of the study by reaching out to 180 users located all over the world.

Meanwhile, a new FCM SeaDataNet Ocean Data View (ODV) data transport format had to be created. As both physical and biological, FCM data were not suitable for the existing standardized SeaDataNet transport formats.

The FCM data management method from the instrument acquisition to the SeaDataNet ingestion is described as follows (see fig.1) : For a project (a cruise for instance), data files acquired by flow cytometry are batch processed, 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 our data centre and the SeaDataNet Request Status Management service (RSM) is made thanks to the Download Manager (SeaDataNet java component tool) which was installed and has been operating since February 27th, 2018.

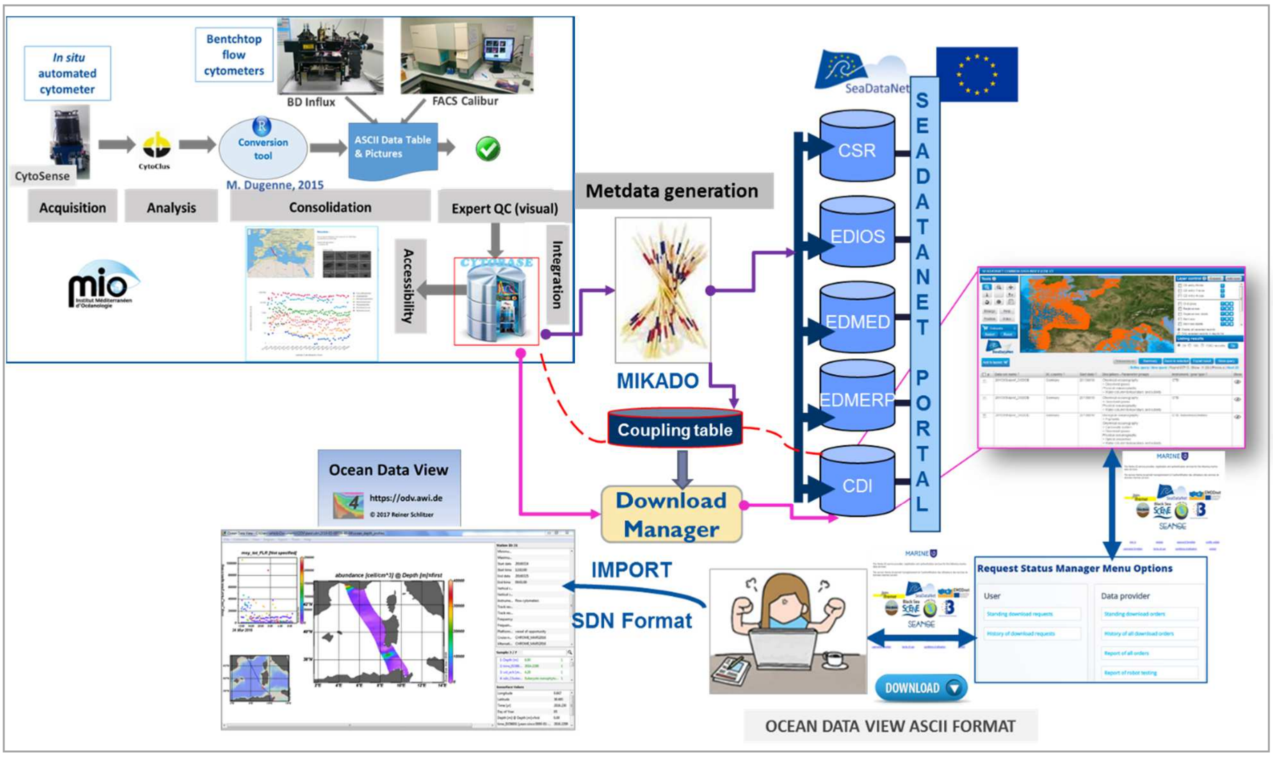

Figure 1: Flow Cytometry (FCM) standardized data management workflow of the CNRS-MIO

Results :

Results of this work consist of 44 new FCM common vocabulary terms that were accepted and integrated into the NERC Vocabulary Server (NVS) operated by the British Oceanographic Data Centre (BODC). Results gathered from the questionnaire allowed us to update some existing vocabularies and add new ones. We keep on ingesting new FCM metadata related to the European Directories (EDMED, EDMERP and CSR) and CDIs which are accessible through SeaDataNet infrastructure (http://seadatanet.maris2.nl/v_cdi_v3/search.asp). The CDI service allows the user to search, browse metadata and request access to data sets via a shopping basket system. Downloaded data are specifically formatted into SeadataNet ODV format for Flow Cytometry data. Every ODV file contains a set of FCM data and has a header (description of the column name (subject), column code (object), unit and used instrument) and a data table (fixed and additional fields). Specific documentation on how to use this format and examples of FCM ODV data and FCM CDI metadata files are available on the SeaDataNet webpage (https://www.seadatanet.org/Standards/Data-Transport-Formats)

Finally, FCM technology and mainly the automated sensors are revolutionizing the biological world by acquiring high resolution and real-time data about the first link of the marine food chain. 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.

Keywords: Flow cytometry datasets, planktonic microbial communities, common vocabulary, data management and interoperability.