The JERICO-S3 project is funded by the European Commission's H2020 Framework Programme under grant agreement No. 871153. Project coordinator: Ifremer, France.

JERICO-S3 DELIVERABLE

Joint European Research Infrastructure for Coastal Observatories
Science, Services, Sustainability

DELIVERABLE #, WP# and full title	JERICO-S3 D6.5 - WP6 - "Best Practices guidelines & strategy for bio-optical sensor and functionality data management"
5 Key words	Best practices, flow cytometry, data management
Lead beneficiary	CEFAS
Lead Author	Véronique Créach
Co-authors	Melilotus Thyssen, Gerald Gregory, Maurice Libes
Contributors	
Final version date/ Submission date (dd.mm.yyyy)	17.07.2024

Nature: **R**
(R = Report, P = Prototype, D = Demonstrator, O = Other)

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

GRANT N°: 871153
PROJECT ACRONYME : JERICO-S3
PROJECT NAME : Joint European Research Infrastructure for Coastal Observatories - Science, services, sustainability
COORDINATOR : Laurent DELAUNEY - Ifremer, France - jerico@ifremer.fr

The JERICO-S3 project is funded by the European Commission's H2020 Framework Programme under grant agreement No. 871153. Project coordinator: Ifremer, France.

DOCUMENT TECHNICAL DESCRIPTION

Document ID	JERICO-S3-WP6-D6.5-17.07.2024-V1.0

REVISION HISTORY

Revision	Date	Modification	Author
V1.0	17.07.2024	Delivery	V. Créach

APPROVALS

	Name	Organisation	Date	Visa
Coordinator	Delauney Laurent	Ifremer	17/07/2024	X
WP Leaders	Peter Thijsse	MARIS	17/07/2024	X

Diffusion list

Consortium beneficiaries	Third parties	Associated Partners	other
X	X		

TABLE of CONTENT

1. Executive Summary **5**
2. Introduction **6**
3. Data Management for Biological Sensors: Case of Flow Cytometry **6**
3.1 Controlled Vocabulary 8
3.2 Metadata Format for FCM 9
3.3 Data Format for FCM 11
3.4 Towards European Infrastructures 12
4. Next Challenges **13**
4.1 Guidance for acquiring Good Data 13
4.2 Guidance for Processing Automatic Gating 15
5. Conclusion **16**
6. References **17**
APPENDIX: ABBREVIATIONS **18**

List of Figures

Figure 1. Schematic of a common flow cytometer, illustrating the fluidic, optical, and electronic systems. Image Source: AAT Bioquest, Inc. 7

Figure 2: Example of alternative names for PFTs and their corresponding BODC IDs. Image source: TT-CYTO workshop. 11

Figure 3: L20- SeaDataNet qualifier flags https://vocab.seadatanet.org/v_bodc_vocab_v2/search.asp?lib=L20. 12

Figure 4: Flow Cytometry data ingestion into SeaDataNet infrastructure using a local database called CYTOBASE at CNRS-MIO. Image Source: TT-CYTO workshop. 13

Figure 5: Theoretical bead concentration vs measured bead concentration. Image Source: TT-CYTO workshop. 14

Figure 6: Bead measurements (FWS total) between 2019 and 2024 to check the stability of the instrument. Image Source: TT-CYTO. 15

Figure 7: Data processing possibilities from raw data to automatic gating process for Cytosense. 16

List of Tables

Table 1: Terms identifying the main FCM groups that appear in the literature and the common nomenclature findable in the https://vocab.nerc.ac.uk/collection/F02/current/. (Thyssen et al., 2022)……..9

Table 2: Mandatory information for FCM metadata to be retained in the organisation…..…11

1. Executive Summary

Novel technologies to study biological components are now frequently used in Operational Ocean Observation. If large amounts of data are now generated, they do not come without challenges in terms of analysis, storage and FAIRness. In the deliverable D6.5, we focused on flow cytometry technology as it is used by several partners in JERICO-RI to generate information on the biodiversity of the phytoplankton communities in the PSSs and IRSs. The deliverable provides what has been achieved in terms of vocabulary, metadata and data format to be stored at the local and European databases as well as a dataflow to SeaDataNet. It also describes progress toward data interoperability by making suggestions for building quality control processes and by reporting data processes recently generated using machine and deep learning approaches.

2. Introduction

The JERICO strategy for optimised use and re-use of observation data from coastal platforms is to develop dataflows from the platforms into the EU data aggregators (CMEMS, EMODnet, SeaDataNet). In order to give access to data to the different users and stakeholders through international aggregators, it is important to develop and apply best practices for innovative technologies for collecting biological data in JERICO-RI. Following the best practices guidelines, the data need to follow FAIR principles (Findable, Accessible, Interoperable and Reusable for humans and machines) which allow the machine-to-machine accessibility and build in real-time or near-time services using EU data aggregators.

The best practices of data management include guidance on metadata and data. They must be adopted as early as possible in the data life cycle. In the last 30 years, a certain number of new sensors for observing and studying the biodiversity in marine environments have been developed. They can be installed on different platforms at sea (e.g. buoys, AUVs, ROVs, Ships of opportunity, FerryBox) and operate autonomously collecting data at high frequency. In contrast to the classical information for biodiversity, most of the data provided by these emerging technologies do not follow the DARWIN Core Archive format (as in use in EuroBIS underpinning EMODnet for biodiversity data) and consequently, they represent a real challenge for European infrastructures. Also, they suffer from a lack of standardisation and quality control processes at the instrument and data management level.

In this deliverable, we focused on a gap in the flow cytometry (FCM) dataflow as this technology is used by a lot of partners in JERICO-RI studying the phytoplankton communities in the Mediterranean Sea, the Adriatic Sea, the Baltic Sea, the North Sea and the Channel. Work on establishing the dataflow of FCM has been initiated in SeaDataCloud. However, the interoperability of the data has never been addressed yet. In this deliverable, we have summarised the previous achievements and the new developments toward the interoperability of the FCM from JERICO-S3. We have also listed the challenges to be overcome before being able to share the FCM with confidence and be part of a digital environment.

3. Data Management for Biological Sensors: Case of Flow Cytometry

Flow cytometry is a laser-based technology that is becoming increasingly popular in plankton ecology. It enables researchers to count individual cells, which can be distinguished based on their specific scatter and fluorescence signatures. Flow cytometers consist of a fluidic system where the water sample containing the cells is jacketed by a fast-moving fluid called sheath fluid in such a way that the cells pass a

laser one by one. Light scatter and fluorescence emitted by the cells are recorded in the detector unit and subsequently processed to characterise the cell (Figure 1).

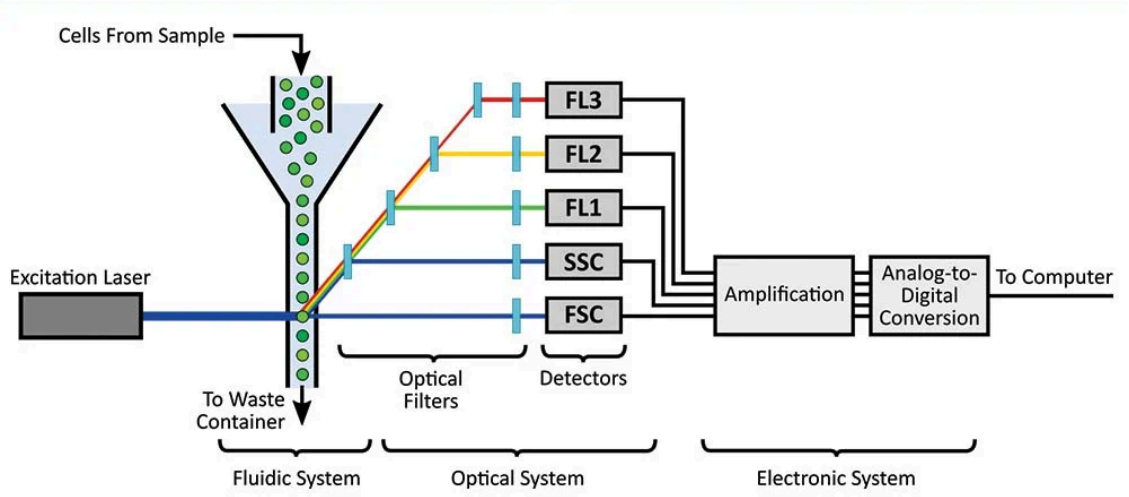

Figure 1. Schematic of a common flow cytometer, illustrating the fluidic, optical, and electronic systems. Image Source: AAT Bioquest, Inc.

Originally being used in laboratory as a lab bench instrument, flow cytometers have been used to assess phytoplankton community in a vessel since early 90' (Yentsch, 1990) dividing the particles into populations that share the same optical properties into three size classes: picophytoplankton (less than 2 µm), nanophytoplankton (between 2 and 20 µm) and microphytoplankton (more than 200 µm). Since the early 2000s, a new generation of instruments have been developed and now can operate autonomously on different platforms such as moorings, attached to a CTD and connected to a FerryBox (e.g. Cytosense, Cytosub, SeaFlow, FlowCytoBot) (Ruiz-Villarereal et al. 2022). They analyse the in-situ phytoplankton communities in real-time and high frequency. They generate a huge amount of data which represents a challenge to process and publish in European databases and archives.
Common and best practices are important to enable users to integrate the information to build and/or update models and products independently from instruments and operators. In the case of flow cytometry, a few initiatives have already provided the first steps towards a data management framework. In SeaDataCloud (10.3030/730960; Grant agreement ID: 730960), preliminary work on a standardised vocabulary and a data pipeline to SeaDataNet (SDN) has been done. More recently, Neeley et al. (2022) have published a technical manual on standards and best practices for reporting flow cytometry observations to enable the validation of future algorithms from new satellites (Ocean color mission Plankton, Aerosol, Cloud, ocean Ecosystem, PACE launched in spring 2024) focusing on phytoplankton growth. However, several issues remain regarding the harmonisation of data processing, and the definition of quality standard procedures which are essential for interoperability.

3.1 Controlled Vocabulary

Flow cytometry does not provide taxonomic-based information on phytoplankton and does not comply with the World Register of Marine Species (WoRMS). However, to support interoperability and reusability a vocabulary needs to be established to remove ambiguity when describing data. An existing vocabulary for flow cytometry was first constructed by SeaDataCloud (Horizon 2020 programme, 730960) and has now been improved and implemented during JERICO-S3. 13 types of phytoplankton (PFTs) are now identified and labeled. They are described in detail in Thyssen et al. (2022). It is registered in the British Oceanographic Data Centre (BODC) using the NERC Vocabulary Server (NVS2.0): http://vocab.nerc.ac.uk/collection/F02/current/. Standardised codes and abbreviations are summarised in Table 1.

Table 1: Terms identifying the main FCM groups that appear in the literature and the common nomenclature findable in the https://vocab.nerc.ac.uk/collection/F02/current/. (Thyssen et al.,2022).

PFTs	BODC ID	name	abbreviation
Picoplankton (2-3 µm)	F0200014	Orange and red fluorescing picophytoplankton Pico cryptophytes	OraPico
	F0200003	Orange fluorescing prokaryote picophytoplankton Syn, Synechococcus, cyanobacteria, PE-type Synechococcus, Crocosphaera, PE-CYAN	OraPicoProk
	F0200013	Red and red only fluorescing picophytoplankton PC-rich Synechococcus, PC-CYAN, PC SYN	RedRedPico
	F0200004	Red only fluorescing picophytoplankton PC-rich Synechococcus, PC-CYAN, PC SYN	RedPico
	F0200002	Red only fluorescing prokaryote picophytoplankton Prochlorophytes, Prochlorococcus, very small red fluorescing bodies, PRO, Prochl	RedPicoProk
Nanophytoplankton (between 2-3 µm and 20 µm)	F0200006	Orange and red fluorescing nanophytoplankton	OraNano

		Cryptomonas, Cryptophytes, HighOrgnano, Nanocyanobacteria, CRYPTO	
	F0200015	Red and red only fluorescing nanophytoplankton Nanocyanobacteria	RedRedNano
	F0200007	Red fluorescing nanophytoplankton with relatively high sideward light scattering properties Coccolithophorids, HighSWSnano, COCCO	HsNano
	F0200005	Red only fluorescing nanophytoplankton Nanoeuk, rednano, Neuk, nanophytoplankton	RedNano
Microphytoplankton (>20 µm)	F0200016	Orange fluorescing microphytoplankton Microphytoplankton	OraMicro
	F0200008	Red only fluorescing microphytoplankton Microphytoplankton	RedMicro

3.2 Metadata Format for FCM

In order to support Reusability metadata should contain all the information needed to describe and explain the data to make the interpretation, the analysis and the processing easier. For all types of data the following information is required:

• Where the data were collected: location (preferably as latitude and longitude) and depth/height
• When the data were collected (date and time in UTC or specified local time zone)
• How the data were collected (e.g. sampling methods, instrument types, analytical techniques)
• How you refer to the data (e.g. station numbers, cast numbers)
• Who collected the data, including name and institution of the data originator(s) and the principal investigator
• What has been done to the data (e.g. details of processing and calibrations applied, algorithms used to compute derived parameters)
• Watch points for other users of the data (e.g. problems encountered and comments on data quality.

To support uptake in SeaDataNet: a general description of an ODV Flow Cytometry (FCM) file for SDN (http://dx.doi.org/10.25607/OBP-568) and a detailed description of ODV fields with CDI metadata and the corresponding data in the ODV FCM format can be found in https://repository.oceanbestpractices.org/handle/11329/1036. In summary, the data table contains 6 fixed fields and a various number of additional fields that require a subject, an object, units, and an instrument. Information on the definition of fields can be found in Kubin (2022).
For local metadata: Table 2 below describes the mandatory elements to record for the FCM sample and should be held in the project information in the organisation as it includes specific information related to the analysis.

Table 2: Mandatory information for FCM metadata to be retained in the organisation.

parameters	Description
project	project acronym name
project starting date	starting date of the project
project ending date	ending date of the project
PI	the person of interest
Cytometer ID	manufacture affiliation/ identification of the machine
station/survey	station/ survey number
File name	Name of the file generated by the flow cytometer
Depth	depth of the sample
Latitude	latitudes coordinates
Longitude	longitude coordinates
Study area	area of sampling
Sample collection	name of the operator during the sampling
Sample analysis	name of the operator during the analysis
Standard reference	beads used as a standard reference
Clustering method	method being used for clustering (manual or automatic)
Observation type	in situ/ex-situ analysis
platform type	sampling platform
platform ID	identification of the sampling platform
Platform nationality	nationality of the platform
Sampling date	sampling date
Analysis date	analysis date is derived from the file name
Trigger channel	the trigger level used to analyse the sample
Trigger level	trigger level in millivolts for the analysis
SWS amplification	Side scatter amplification in millivolts for the analysis (PMT)

FLO amplification	Orange fluorescence amplification in millivolts for the analysis (PMT)
FLR amplification	Red fluorescence amplification in millivolts for the analysis (PMT)

3.3 Data Format for FCM

This paragraph only concerns the level 2 submission data which means data that has been processed by an initial round of analysis or computation to clean the data and assess basic quality measures. Only the concentration of particles per PFTs (particles per µl) must be reported in SDN as well as the volume analysed (m3). The names given by the experts during the manual clustering of the PFTs can be different from the BODC Vocabulary IDs. These alternative names can be stored in the local database. However, they will need to translate to BODC Vocabulary IDs when the concentrations are reported in SDN (Figure 2). Size and fluorescence are arbitrary values and differ for each flow cytometer. Consequently, these values need to be calibrated with beads or algae for size, and a discrete sample of chlorophyll-*a* analysed by HPLC or spectrophotometry for fluorescence. As no best practices and quality control processes have been discussed and agreed upon, these variables should be only stored in the local database.

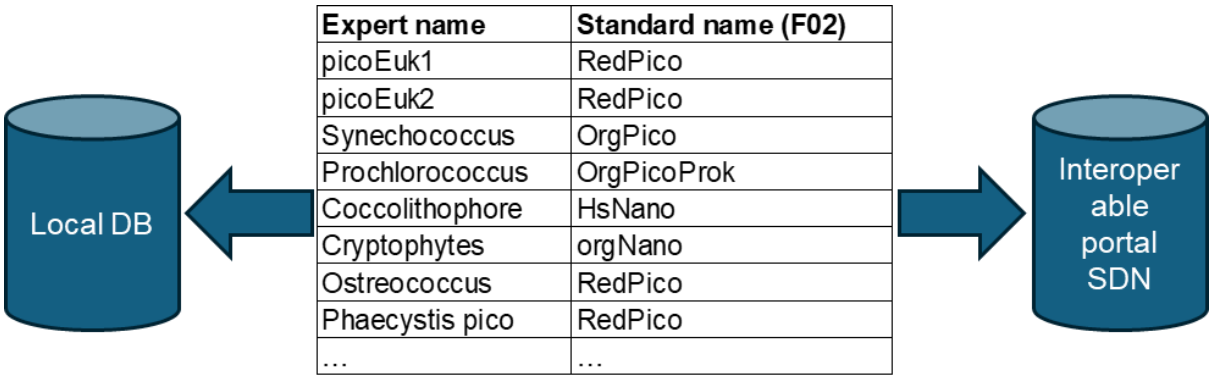

Figure 2: Example of alternative names for PFTs and their corresponding BODC IDs. Image source: TT-CYTO workshop.

Data quality is a measure of a data set's condition based on factors such as accuracy, completeness, consistency, reliability and validity. Data quality assessment aims to detect and as appropriate flag for review or remove such abnormalities in measurements that likely do not possess any underlying biological causes, and thus are likely caused by technical errors. Technical issues resulting from instrumental or procedural variations during acquisition can bias the statistics of the obtained cell subpopulation and can impact the quality of the cytometry data and the subsequent analysis results. Clogs can result in abrupt changes in the fluorescence in the time domain analysis. Other issues such as unstable data acquisition can result in a shift in the means of the populations analysed, which can pose challenges for gating.

These data should be identified or potentially removed by the user, either automatically or manually, before being passed to the downstream analysis. Figure 3 shows the code quality and its descriptions for SDN. Without quality control processes in place yet, the FCM data must be validated by the scientist or operator before being stored and published in SDN. Consequently, the data present in SDN are all qualified with flag "1" in the ODV file meaning "good value".

Flag Description	ODV	GTSPP	ARGO	SEADATANET	ESEAS	WOD	WODSTATION	WOCEBOTTLE	WOCECTD	WOCESAMPLE	QARTOD	BODC	PANGAEA	SMHI	OceanSITES	IODE
no quality control	1	0	0	**0**	0	0	0	2	2	2	0	Q	*	blank	0	2
good value	0	1	1	**1**	1	0	0	2	2	2	3	blank	blank	blank	1	1
probably good value	0	2	2	**2**	1	0	0	2	2	2	3	blank	blank	blank	2	1
probably bad value	4	3	3	**3**	3	4	3	3	3	7	2	K	?	?	3	3
bad value	8	4	4	**4**	4	4	3	4	4	7	1	K	/	B	4	4
changed value	1	5	5	**5**	2	0	0	2	2	2	0	R	*	|	5	2
value below detection	1	0	0	**6**	0	0	0	2	2	2	0	<	<	<	0	2
value in excess	1	0	0	**7**	0	0	0	2	2	2	0	>	>	>	0	2
interpolated value	1	0	8	**8**	2	0	0	2	2	2	0	T	*	|	0	2
missing value	1	9	9	**9**	9	0	0	5	5	5	9	N	*	B	9	9
value phenomenon uncertain	1	0	0	**A**	0	0	0	2	2	2	0	Q	*	B	0	2

Figure 3: L20- SeaDataNet qualifier flags https://vocab.seadatanet.org/v_bodc_vocab_v2/search.asp?lib=L20

3.4 Towards European Infrastructures

The FCM data do not follow the Darwin Core Archive format as the PFTs do not have a taxonomy reference. Consequently, the FCM data cannot be directly registered in EMODnet biology. SDN represents the best alternative and already has 24 records of FCM data from the marine environment. These records have been either processed by one of the IODE National Oceanographic Data Centres (NODCs) or at the local level with data management support such as the CYTOBASE now established at CNRS-MIO (Figure 4). Recently, technical developments were put forward in the EMODnet Biology portal to allow a wider community to search for FCM data from SDN. This new development gives more visibility to the FCM data and increases the possibility of sharing the information with other international networks such as EUROBIS/OBIS.

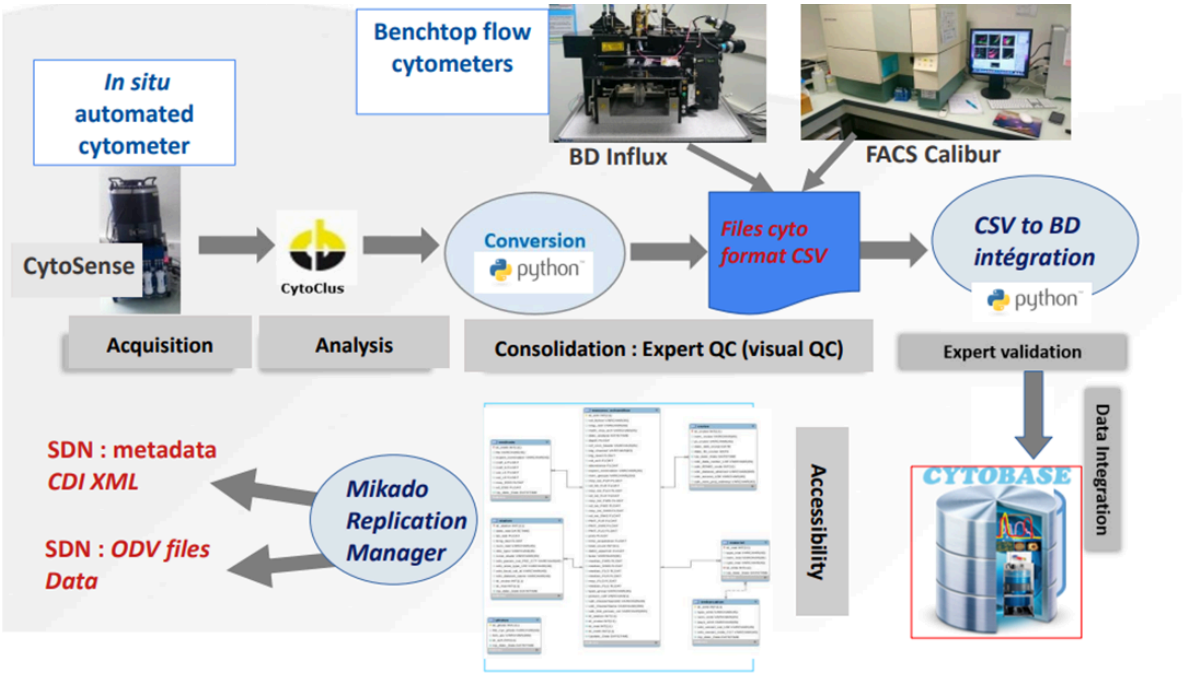

Figure 4: Flow Cytometry data ingestion into SeaDataNet infrastructure using a local database called CYTOBASE at CNRS-MIO. Image Source: Workshop TT-CYTO workshop.

4. Next Challenges

QC data is mandatory for interoperability and reusability. Besides the definition of a standard vocabulary for the different phytoplankton functional types, harmonised quality control processes for FCM data from the marine environment does not exist yet. During a workshop "TT-CYTO: Tips & Tricks towards Flow Cytometry Data FAIRness in June 2024 (Wimereux, France; co-funded by Euromarine, and EU projects JERICO-S3 and OBAMA-NEXT, and IFSEA) quality control procedures were presented and discussed to prepare a best practice document.

4.1 Guidance for acquiring Good Data

The procedures below are known and followed by the users. They define the limits of the instrument for optimising protocols for measuring phytoplankton communities.
Avoid coincidence and define the best parameters for the measurement. Coincidence in flow cytometry refers to the situation where more than one cell is in the laser path at the same time. Coincidence underestimates the number of cells of interest and consequently the concentration of the PFTs. The rate of coincidence is determined by comparing a theoretical and measured number of beads at different speeds (Figure 5). A variation of 10% is acceptable.

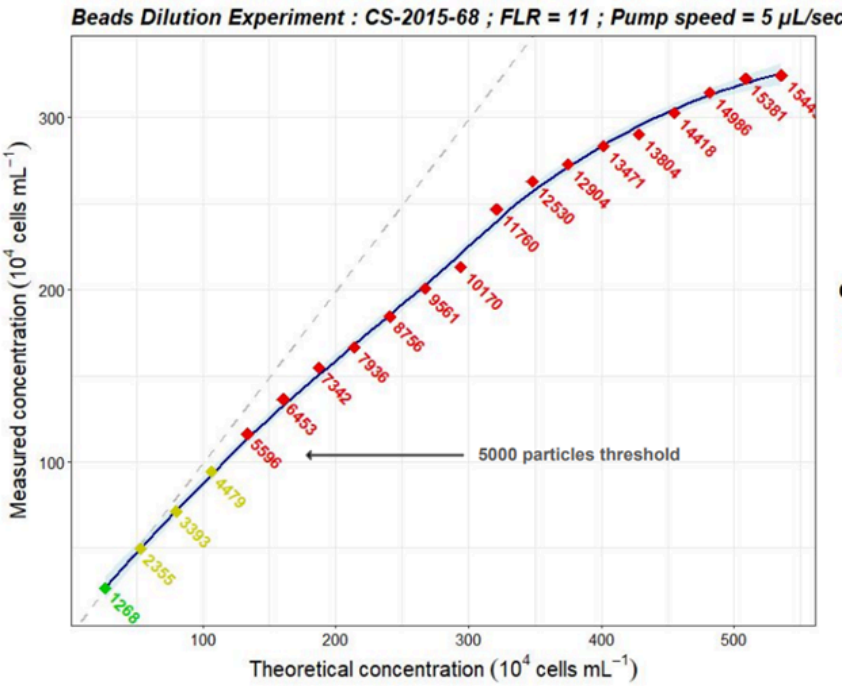

Figure 5: Theoretical bead concentration vs measured bead concentration. Image Source: TT-CYTO workshop.

Get a good representativity of the phytoplankton community, by decreasing the trigger level enough to be able to measure the lowest size and fluorescence phytoplankton functional types. At the same time, it is important to avoid capturing too much electronic noise from the instrument because it will increase the volume of files and decrease the representativity of the phytoplankton measured in the sample. The instrument noise should be no more than 10% of the total particles measured with the flow cytometer. Also, the analysis volume depends on the number of particles in the PFTs. For a PFT with a low concentration of particles, it will be necessary to analyse a high volume.

Test the performance of the instrument (Figure 6) but also to help with the determination of the PFTs and determine the size of the particles, beads should be added in samples or blanks.

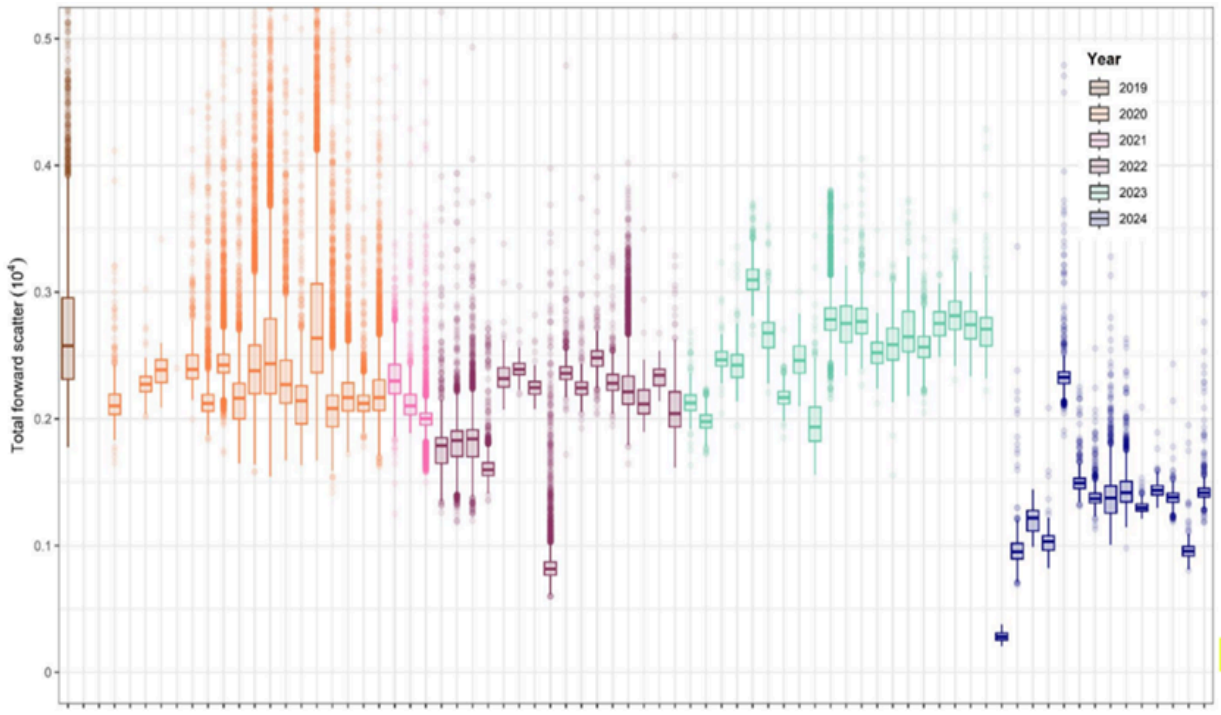

Figure 6: Bead measurements (total FWS) between 2019 and 2024 to check the stability of the instrument. Image Source: TT-CYTO.

4.2 Guidance for Processing Automatic Gating

Gating is an inherent component of FCM data analysis. It divides the particles according to their physical (Forward Scatter, Side Scatter) and fluorescent characteristics (Red fluorescence, Orange fluorescence). Manual gating is still the most common practice in FCM for environmental studies. However, this way of defining arbitrary groups is not always objective and can lead to errors, when clusters overlap, shift positions or content a very limited number of particles. Between the first automatic gating applied to algae by Balfoot et al. (1992) and the most recent deep learning approach by Fuchs et al. (2022), several initiatives based on statistical methods (Wacquet et al., 2013), on Random Forests (Thomas et al., 2018; Schmidt et al., 2020) have been tested and published. If the previous initiatives needed clustering software as an intermediate step (https://github.com/RobeeF/phyto_curves_reco.git), today the data generated by the instrument can be directly extracted using an API (https://github.com/Cytobuoy/CyzFile-API) and processed after being transformed as JSON file (https://github.com/OBAMANEXT/cyz2json). This new data flow (Figure 7) has been already used to build a Random Forests approach which gives around 97% similarity with the manual clustering from the experts (Lanoy, 2024).

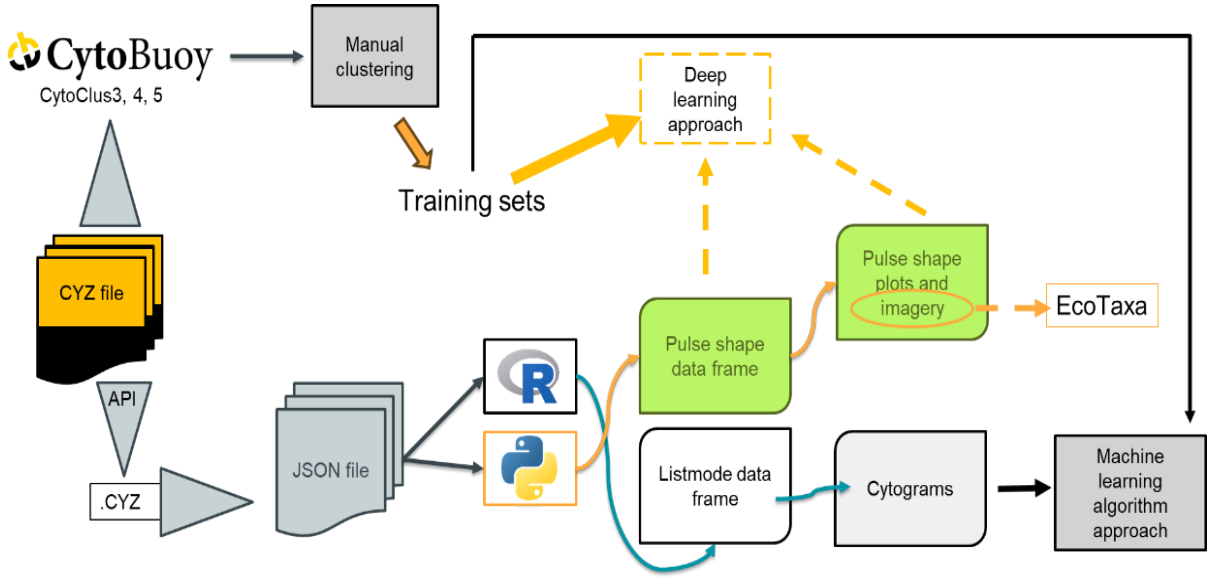

Figure 7: Data processing possibilities from raw data to automatic gating process for Cytosense.

5. Conclusion

During JERICO-S3, a step forward towards FAIR data for flow cytometry has been made by publishing an agreed vocabulary for marine microbes and establishing a new dataflow that allows the development of machine learning and deep learning approaches. Metadata and data format previously defined have been discussed and implemented. Collaboration with instrument makers has been essential to get access to all information related to the data and building interface as well as the collaboration with data scientists. If a control quality process needed for interoperability is not yet implemented, best practices have been shared between experts and considered for publication in Ocean Best Practices. However, challenges remain. Machine learning and deep learning processes need consensual training sets that can only be built manually and consequently depend on the expertise of the scientists. Also, facilities for data archives, local databases, and expertise in formatting processes towards European Infrastructures need to be available at the organisation level. Solutions are available but they are costly and impossible to afford at the project level. JERICO-RI is an exceptional platform for scientists to exchange expertise but also represents a great opportunity to build and offer the facilities needed (e.g. control quality processes, formatting facilities) for the FAIRness of the data from sensor-based biodiversity observations.

6. References

- Balfoort, H., Snoek, J., Smiths, J., Breedveld, L., Hofstraat, J., & Ringelberg, J. (1992). Automatic identification of algae: neural network analysis of flow cytometric data. Journal of Plankton Research,14(4), 575-589. doi:10.1093/plankt/14.4.575.
- Fuchs, R., Thyssen, M., Creach, V., Dugenne, M., Izard, L., Latimier, M., Louchart,A., Marrec, P., Rijkeboer, M., Grégori, G., & Pommeret, D. (2022). Automatic recognition of flow cytometric phytoplankton functional groups using convolutional neural networks. Limnology and Oceanography: Methods, 20. Doi:10.1002/lom3.10493.
- Kubin, E. (2022). Updated guidelines for SeaDataNet ODV production: Eutrophication & Contaminants. doi:10.13120/C1933032-9FA9-4678-8539-EFFA1560921C.
- Lanoy, L. (2024). Rapport de stage for Master 2nd year. June 2024. pp24.
- Neeley, M., Sourisseau, M., Anderson, P., Cusack C. et al. (2022). Novel methodologies for providing in situ Data to HAB Early Warning Systems in the European Atlantic Area: The PRIMROSE Experience. Frontiers in Marine Sciences, 9. doi:10.3389/fmars.2022.791329.
- Schmidt, K. C., S. L. Jackrel, D. J. Smith, G. J. Dick, & V. J. Denef. (2020). Genotype and host microbiome alter competitive interactions between microcystis aeruginosa and chlorella sorokiniana. Harmful Algae, 9: 101939. doi:10.1016/j.hal.2020.101939.
- Thomas, M. K., Fontana, S., Reyes, M., & Pomati, F. (2018). Quantifying cell densities and biovolumes of phytoplankton communities and functional groups using scanning flow cytometry, machine learning and unsupervised clustering. PLOS ONE,13(5). doi:10.1371/journal.pone.0196225.
- Thyssen, M., Grégori, G., Créach, V., Lahbib, S., et al. (2022). Interoperable vocabulary for marine microbial flow cytometry. Frontiers in Marine Science,9. doi:10.1093/plankt/fbn005.
- Wacquet, G., Poisson Caillault, í., Hamad, D., & HéBert, P. (2013). Constrained spectral embedding for K-way data clustering. Pattern Recognition Letters,34(9), 1009-1017. doi:101016/j.patrec.2013.02.003.
- Yentsch, C.M. (1990). Environmental health: flow cytometric methods to assess our water world. Methods Cell Biol., 33. doi: 10.1016/s0091-679x(08)60555-5.

APPENDIX: ABBREVIATIONS

AUVs: Autonomous Underwater Vehicles
CMEMS: Copernicus Marine Environment Monitoring Service
EMODNet: European Marine Observation and Data Network
EuroOBIS: European Node of the International Ocean Biodiversity Information
FAIR: Findable Accessible Interoperable Reusable
FCM: Flow Cytometry
JERICO-RI: JERICO Research Infrastructure
IRSs: Integrated Regional sites
PFTs: Phytoplankton Functional Types
PSs: Pilot supersites
NODC: National Oceanographic Data Center
QC: Quality Control
ROVs: Remotely Operated Underwater Vehicles
SDN: SeaDataNet