SeaDataCloud

CNRS data flow to SeaDataNet

- *François Gaudin, Fabrice Mendes, Pascal Calvat, Yolanda del Amo* – OASU, UMS POREA 2567, UMR EPOC 5805 - Université de Bordeaux
- *Soumaya Lahbib, Maurice Libes, LLoyd Izard, Melilotus Thyssen, Gérald Grégori* - MIO UMR 7294 , OSU Pytheas UMS 3470 - Université Aix Marseille
- *Mark Hoebeke, Fabienne Rigaut, Nathalie Simon* - Station Biologique de Roscoff, CNRS - Sorbonne Université

Institut Pythéas
Observatoire des Sciences de l'Univers
Aix+Marseille Université

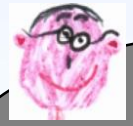
mio Institut Méditerranéen d'Océanologie

Observatoire Aquitain des Sciences de l'Univers

EPOC

CNRS • SORBONNE UNIVERSITÉ
Station Biologique de Roscoff

3 (out of 27) Observatories of the Science of the Universe (OSU)

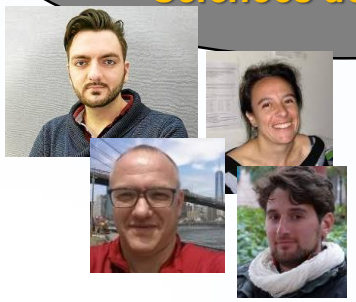


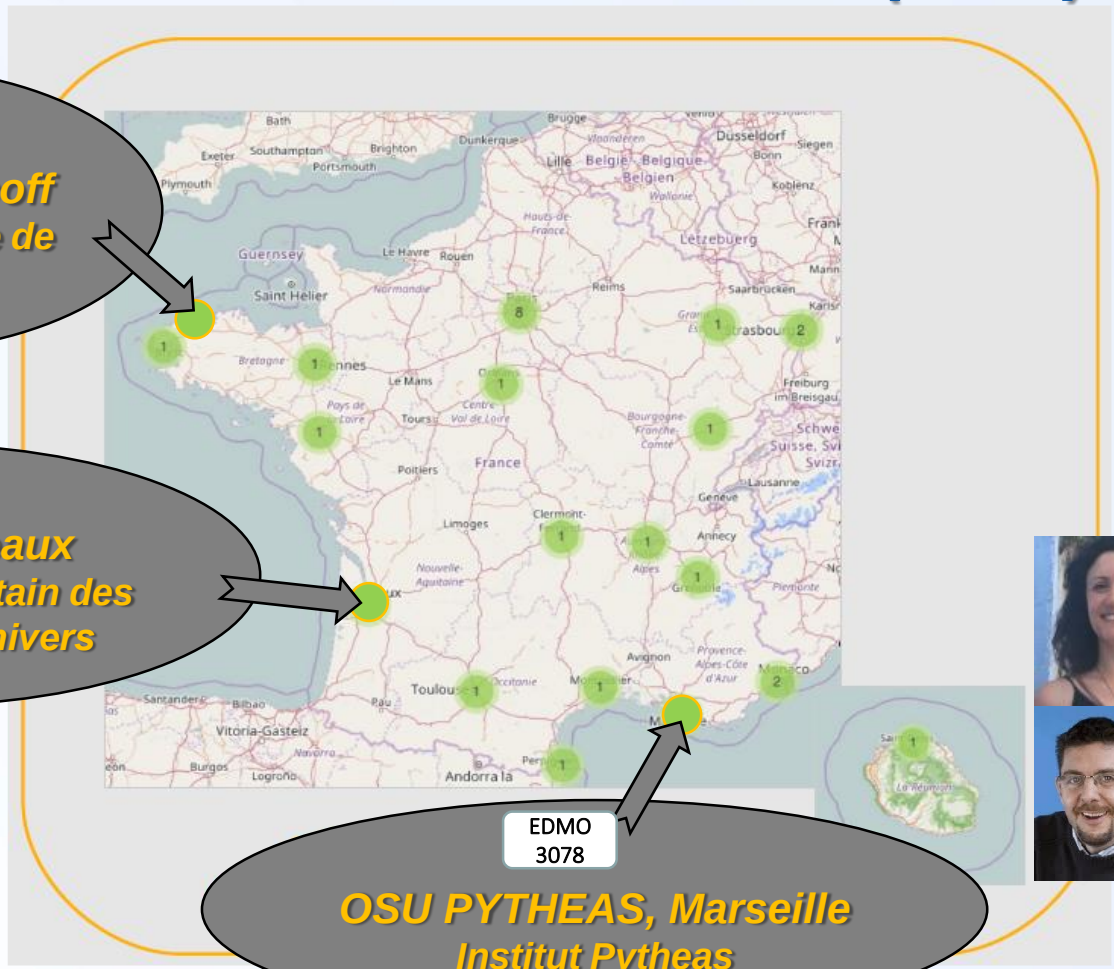
EDMO 521

OSU-SBR, Roscoff
Station Biologique de Roscoff

EDMO 1002

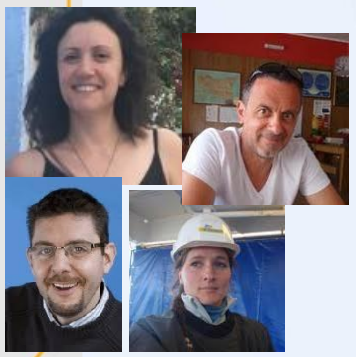
OASU, Bordeaux
Observatoire Aquitain des Sciences de l'Univers





EDMO 3078

OSU PYTHEAS, Marseille
Institut Pytheas



The National Observation Services (NOS)

French observation networks structuration within the Research Infrastructure for Coastal Ocean and Seashore ILICO

The datasets & the databases

CNRS • SORBONNE UNIVERSITÉ
Station Biologique de Roscoff

resomar

PELAGOS

Observatoire Aquitain des Sciences de l'Univers

EPOC

somlit
Service d'Observation en Milieu Littoral

Institut Pythéas
Observatoire des Sciences de l'Univers
Aix*Marseille Université

mio
Institut Méditerranéen d'Océanologie

CYTOBASE

SEADATANET PORTAL

Flow Cytometry (FCM) data - CYTOBASE

Context & Initial aims

- Flow cytometry expert team at MIO (Melilotus Thyssen and Gerald Gregori)
- Many scientific campaigns and projects *(67 CDIs)*
- Abundant and complex data treatment

- Needs (objectives)
  - efficient software tools & workflow to manage data
  - interoperability for sharing data

SeaDataCloud WP9.2.5

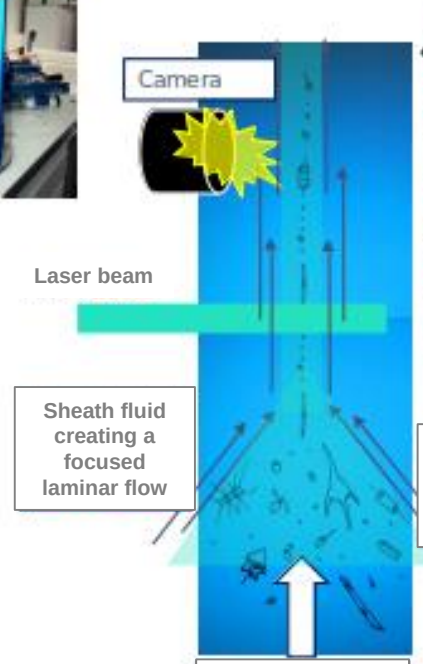
"Ingesting, validating, long-term storage and access of Flow Cytometry data"

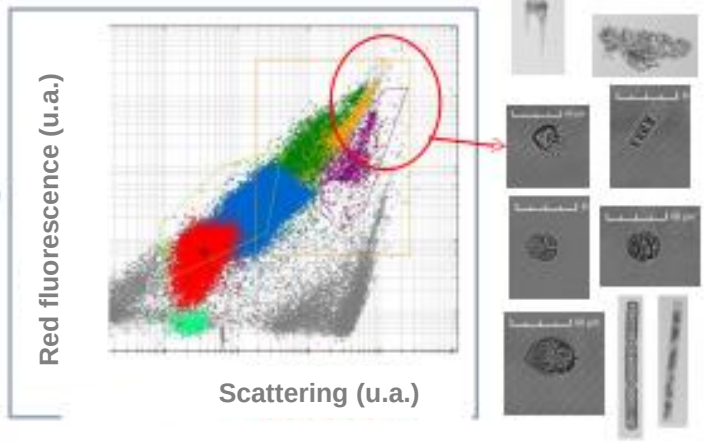
1. FCM Common Vocabulary
2. Data Transport Format for FCM data
3. Ingestion into SDN Infrastructure

Institut Pythéas
Observatoire des Sciences de l'Univers
Aix+Marseille Université

mio Institut Méditerranéen d'Océanologie

Flow cytometry (FCM) for marine research





Cytogram or picture describing the phytoplankton community
Up to 7 "functional" groups :
Prochlorococcus
Synechococcus
Picoeukaryotes
Nanoeukaryotes
Microphytoplankton
Cryptophytes like
Coccolitophorideae like

Institut Pythéas
Observatoire des Sciences de l'Univers
Aix*Marseille Université

mio
Institut Méditerranéen d'Océanologie

SeaDataCloud

Flow cytometry (FCM) for marine research

- nom_cruise :
- num_stat :
- sampling_date, analysis_date
- longitude / latitude
- sdn_local_cdi_id : FA35102016_CHROME_OCT_2016_FCMW
- sdn_edmo_code : 3078
- COL_instrument : tool1209
- bot_depth + depth :
- vol_ech :
- sdn_ClusterName : *Coccolithophores*
- sdn_ClusterNameID : *SDN:F02::F0200007*
- abundance : 19.4796
- Optical properties : *Red Fluorescence, Orange Fluorescence, Forward Scatter, Side Scatter*
  - *moy_tot_FLR QV_moy_tot_FLR sd_tot_FLR QV_sd_tot_FLR*
  - *moy_tot_FLO QV_moy_tot_FLO sd_tot_FLO QV_sd_tot_FLO*
  - *moy_tot_FWS QV_moy_tot_FWS sd_tot_FWS QV_sd_tot_FWS*
  - *moy_tot_SWS QV_moy_tot_SWS sd_tot_SWS QV_sd_tot_SWS*

Institut Pythéas
Observatoire des Sciences de l'Univers
Aix*Marseille Université

mio Institut Méditerranéen d'Océanologie

Flow Cytometry (FCM) data - CYTOBASE

For SDC tasks…

- Need to manage *in situ* real time flow cytometry data
- Converting FCM data files into CSV readable format
  (++ Python quality controls)
- Adding metadata

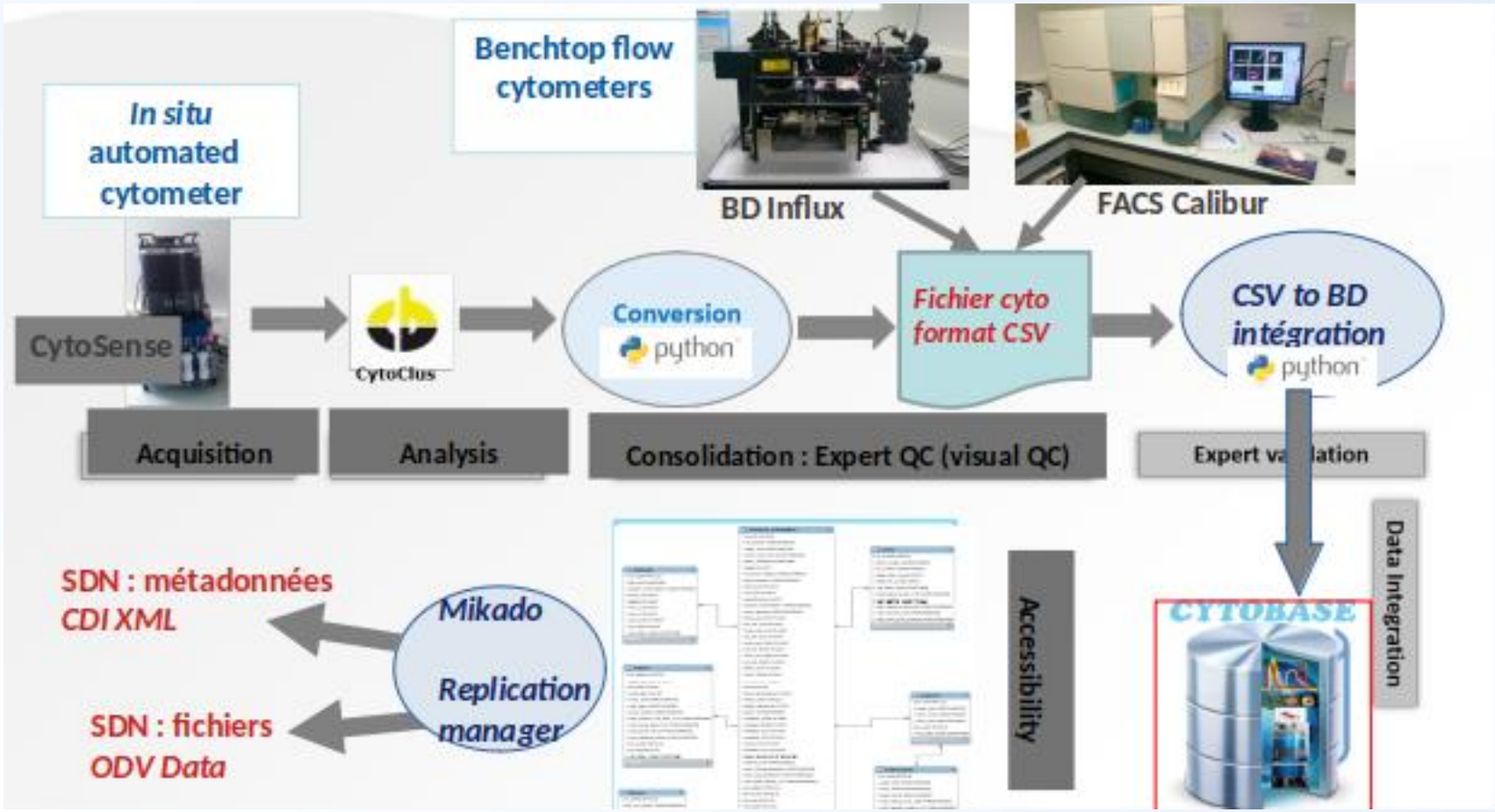
- Producing a common standard FCM voc.

- Inserting FCM data into a DB in order to perform queries
  (++ Python quality controls)

➔ Sharing FCM data with scientific community
(with metadata, with QC, in interoperable formats)

SeaDataCloud

Flow Cytometry data workflow



Institut Pythéas
Observatoire des Sciences de l'Univers
Aix+Marseille Université

mio
Institut Méditerranéen d'Océanologie

Microphytoplankton biodiversity data - PELAGOS DB

- specie's names
- abundances

Context & Initial aims

- Publishing biodiversity datasets *(2 CDIs)*
- Long-term coastal phytoplankton biodiversity data from NOS
- 24 measurements / year, since April 2000
- Time-consuming analysis through optical microscopy by experts
- Storage in the national PELAGOS DB hosted at Roscoff Marine Station

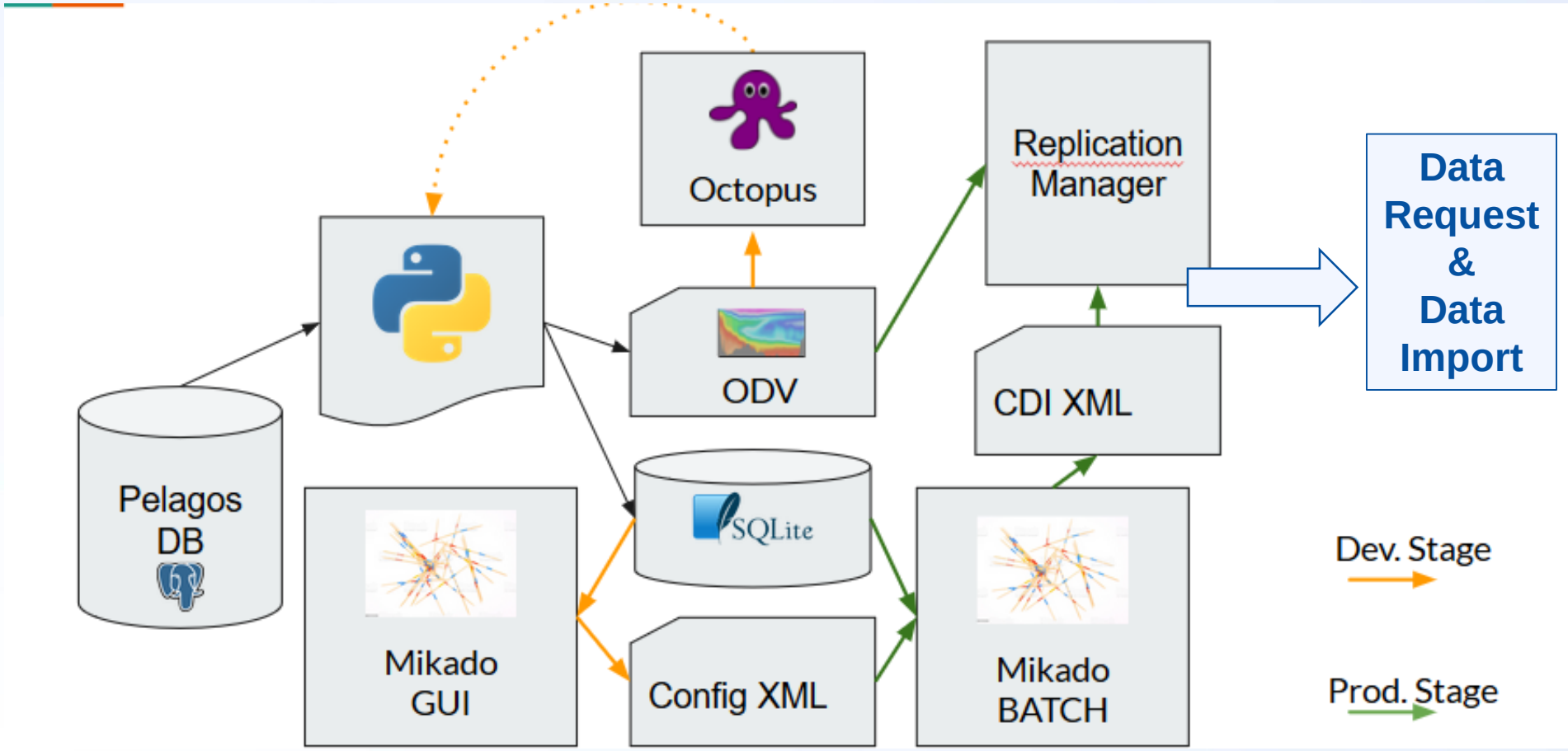
Expected Main Challenges

- Mapping PELAGOS data models with SeaDataNet
- Automating the workflow, from data extraction to data publishing

Expected Minor Challenges

- Setting up SDN technical components
- Mastering SDN tools

Microphytoplankton biodiversity data workflow



Physical, chemical, biological data – SOMLIT NOS DB

Context & Initial aims

- Bimonthly sampling from 1996
- 16 Essential Ocean Variables & pico-nano-plankton
- 12 ecosystems
- 3 main data types
  - ❖ hydro-biol. time-series *(50 CDIs)*
  - ❖ CTD *(3913 CDIs)*
  - ❖ pico-nano-plankton – FCM *(20 CDIs)*
- QC checks

But

- Metadata missing in DB
- Human approbation needed for data distribution
- File format not official standards

Challenges

- Integrating all metadata (contacting every expert in the 12 stations)
- Discovering & Learning SDN ecosystem
- Mapping data models

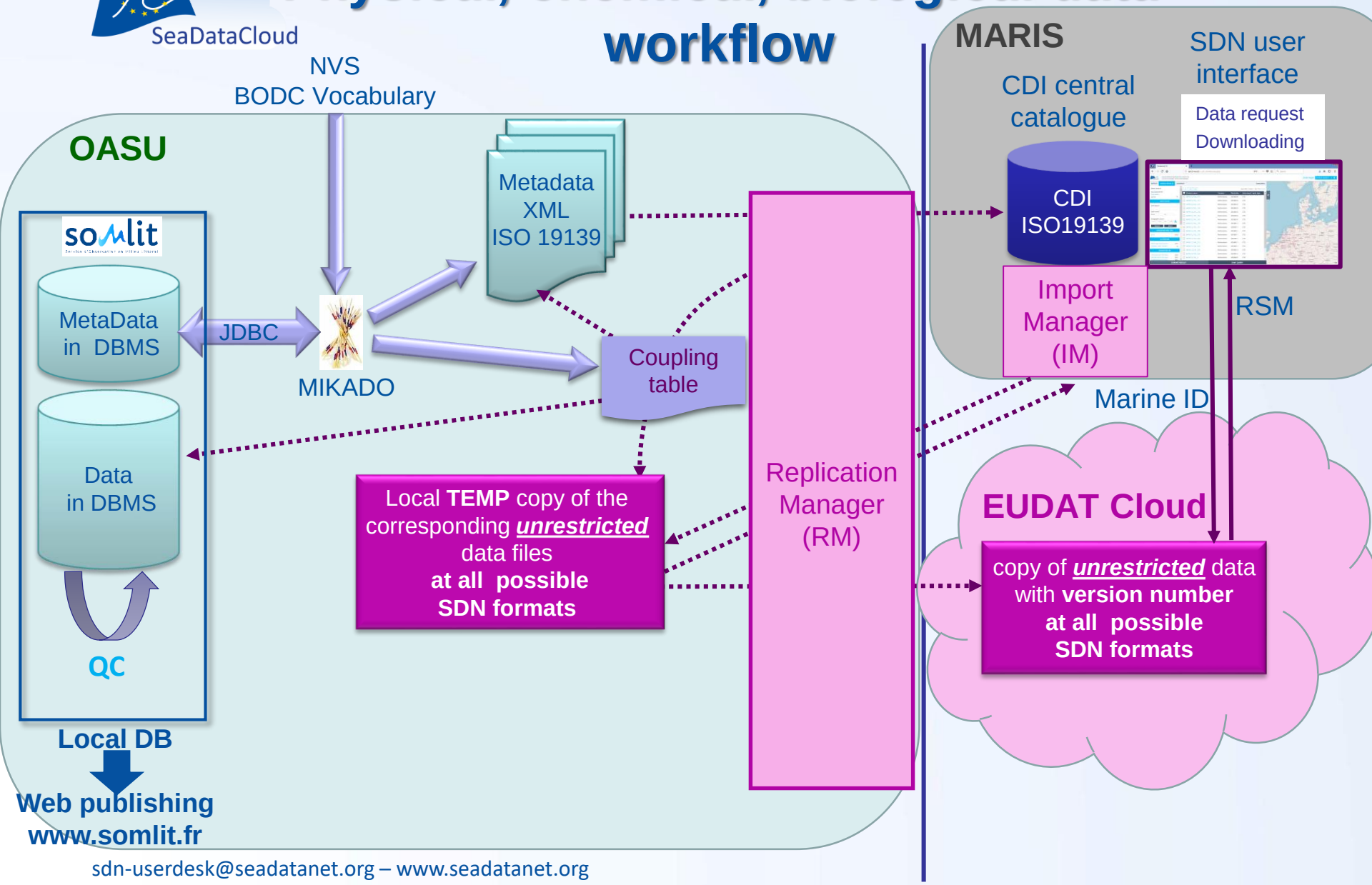
Main Goal

- International DB publishing
- International standards

Observatoire Aquitain des Sciences de l'Univers

EPOC

Physical, chemical, biological data workflow



Feedback statement – PROs & CONs

PROs

- ✓ SDN allows the ***distribution*** of data in standard formats through a nice and ***intuitive*** interface : datasets available in SDN (no more sending mails with attachments)
- ✓ ***European visibility*** of flow cytometry and SOMLIT data, laboratories, projects and originators (marine stations and labs)
- ✓ ***Interoperability*** with ***standardization*** of data file formats, and parameter names
- ✓ Gives a ***framework*** for data management in a standardized way
- ✓ Work in harmony and in ***accordance*** with the ***requirements of national datapoles (e.g. Odatis)***
- ✓ Definition of a ***standardized BODC controlled vocabulary*** among cytometrists
- ✓ ***Excellent help (& THANKS !!!)***
  - From collating centre (IFREMER), SDN userdesk, CDI-help desk
  - Training sessions (general and personal)
  - Documentation
  - "Informal" mail exchange with knowledgeable insiders

Feedback statement – PROs & CONs

PROs

- ✓ SDN allows the ***distribution*** of data in standard formats through a nice and ***intuitive*** interface : datasets available in SDN (no more sending mails with attachments)
- ✓ ***European visibility*** of flow cytometry and SOMLIT data, laboratories, projects and originators (marine stations and labs)
- ✓ ***Interoperability*** with ***standardization*** of data file formats, and parameter names
- ✓ Gives a ***framework*** for data management in a standardized way
- ✓ Work in harmony and in ***accordance*** with the ***requirements of national datapoles (e.g. Odatis)***
- ✓ Definition of a ***standardized BODC controlled vocabulary*** among cytometrists
- ✓ ***Excellent help (& THANKS !!!)***
  - From collating centre (IFREMER), SDN userdesk, CDI-help desk
  - Training sessions (general and personal)
  - Documentation
  - "Informal" mail exchange with knowledgeable insiders

DIFFICULTIES/ CONs

- ✓ Taming the ***complexity*** of SDN workflow, P0n thesaurus and tools (e.g. Mikado, Nemo, RM…)
- ✓ ***Understanding*** coupling tables and mapping files format
- ✓ Finding a ***compromise*** solution for the ***detail level*** of metadata from long-term series (past analytical/sampling methods) and its feasibility (especially in order not to "cut" series)
- ✓ ***Length*** of chain for processing ***changes*** (*e.g.* updating EDIOS or C17 vocabulary list)
- ✓ Metadata control and mapping by research experts was a big task
- ✓ Re-definition of ***dataset notion*** to match SDNet's : ***re-structuration local DB***
- ✓ ***Underestimation*** of the developer tasks for matching SDN needs

Feedback statement

NEW METHODS of management ?

- ✓ Look out and ***track updates*** of BODC Vocs, of European recommendations for standards
- ✓ Use of ***thesaurus*** and ***standard file formats*** (*e.g.* odv, NetCDF…)
- ✓ Addition of ***QC_Flags*** to some data types (*e.g*. CTD)
- ✓ Better ***traceability*** of data modifications (*e.g*. corrections, adds…)
- ✓ Local and new reflection on ***« dataset » definition***

Issues to keep in mind

- ✓ Data producers need to be (made) aware of the importance of ***describing*** their datasets using ***metadata*** as early as possible, and can still be reluctant to participate in this task.
- ✓ There was a significant ***effort*** to achieve mapping between the in-house data models and the SDN data format… But it should ***pay off*** over time as the surveys are standardized.
- ✓ (bio)ODV as we understand it may not be the best suited data format for storing “rich” biodiversity datasets
- ✓ Small improvement : the ***reason/objectives*** for what users download the data is not displayed in the Import Manager

SeaDataCloud

Thank you,

hoping to keep ingesting new data…

- *François Gaudin, Fabrice Mendes, Pascal Calvat, Yolanda del Amo* – OASU, UMS POREA 2567, UMR EPOC 5805 - Université de Bordeaux
- *Soumaya Lahbib, Maurice Libes, LLoyd Izard, Melilotus Thyssen, Gérald Grégori* - MIO UMR 7294 , OSU Pytheas UMS 3470 - Université Aix Marseille
- *Mark Hoebeke, Fabienne Rigaut, Nathalie Simon* - Station Biologique de Roscoff, CNRS - Sorbonne Université

Institut Pythéas Observatoire des Sciences de l'Univers Aix+Marseille Université

Observatoire Aquitain des Sciences de l'Univers

EPOC

CNRS • SORBONNE UNIVERSITÉ Station Biologique de Roscoff

mio Institut Méditerranéen d'Océanologie

université de BORDEAUX

SORBONNE UNIVERSITÉ