SeaDataNet

PAN-EUROPEAN INFRASTRUCTURE FOR OCEAN & MARINE DATA MANAGEMENT

WP10
Development of Products

S.Simoncelli (INGV, Italy) and WP10 Regional Coordinators

SeaDataNet

OUTLINE

1. WP10 Objectives/Role of Partners/Deliverables
2. Review of first year activities
3. Collaboration with MyOcean Project
4. Aggregated Data Sets→ aggregation procedure
5. Quality Assessment procedure
6. Analysis of data anomalies
7. Conclusions
8. Future Work Plan

WP10 Objectives

1. To **validate data access** and processing services provided by SDN infrastructure
2. To analyze the coherency, coverage and quality of the data sets at full basin scale for European seas
3. To generate **data products** using SDN historical database
4. To create regional **aggregated data sets**
5. To compute **statistical products** (climatologies) from data collections
6. To use ODV and DIVA tools for QC and climatology computation
7. To coordinate and support the harmonization of products between sea basins
8. To make products and relative documentation available to the users

Role of Partners

- Work done by regional groups. Regional coordinators (RC) lead the activities
- A leading team (RC + data centers) guide the QC process
- Activities done in communication with ongoing projects (MyOcean) in order to develop true synergies at regional level

Med Sea	INGV	*Simona Simoncelli*
Black Sea	METU	*Devrim Tezcan*
North Atlantic	IFREMER	*Christine Coatanoan*
North Sea	MUMM	*Serge Scory*
Arctic Waters	IMR	*Helge Sagen*
Baltic Sea	SMHI	*Bäck Örjan*

WP10 Deliverables

Common specifications, selected parameters, definition of QC, aggregation and analysis procedures — **Sept 2012**

V1
AGGREGATED DATASET — **Sept 2013**

Examples of data products
gridded fields/ climatological profiles and relative std — **Sept 2014**

V2
AGGREGATED DATASET — **Sept 2015**

Delive-rable Number [61]	Deliverable Title	Delivery date [64]
D10.1	Common specifications, selected parameters, aggregation, control and processing procedures	12
D10.2	First release of the aggregated data sets products	24
D10.3	Release of examples of data products	36
D10.4	Final version of aggregated data sets	46

SeaDataNet

First Year Outcome

1. General overview of SeaDataNet infrastructure content per Sea Region;
2. Common Specifications for the regional aggregated data sets and the statistical products
3. Definition of **Aggregation Procedure**
4. Definition of **Quality Control Procedure**
5. Start of **SDN-MyOcean INSTAC collaboration** to elaborate jointly Regional T&S product

PAN-EUROPEAN INFRASTRUCTURE FOR OCEAN & MARINE DATA MANAGEMENT

SeaDataNet

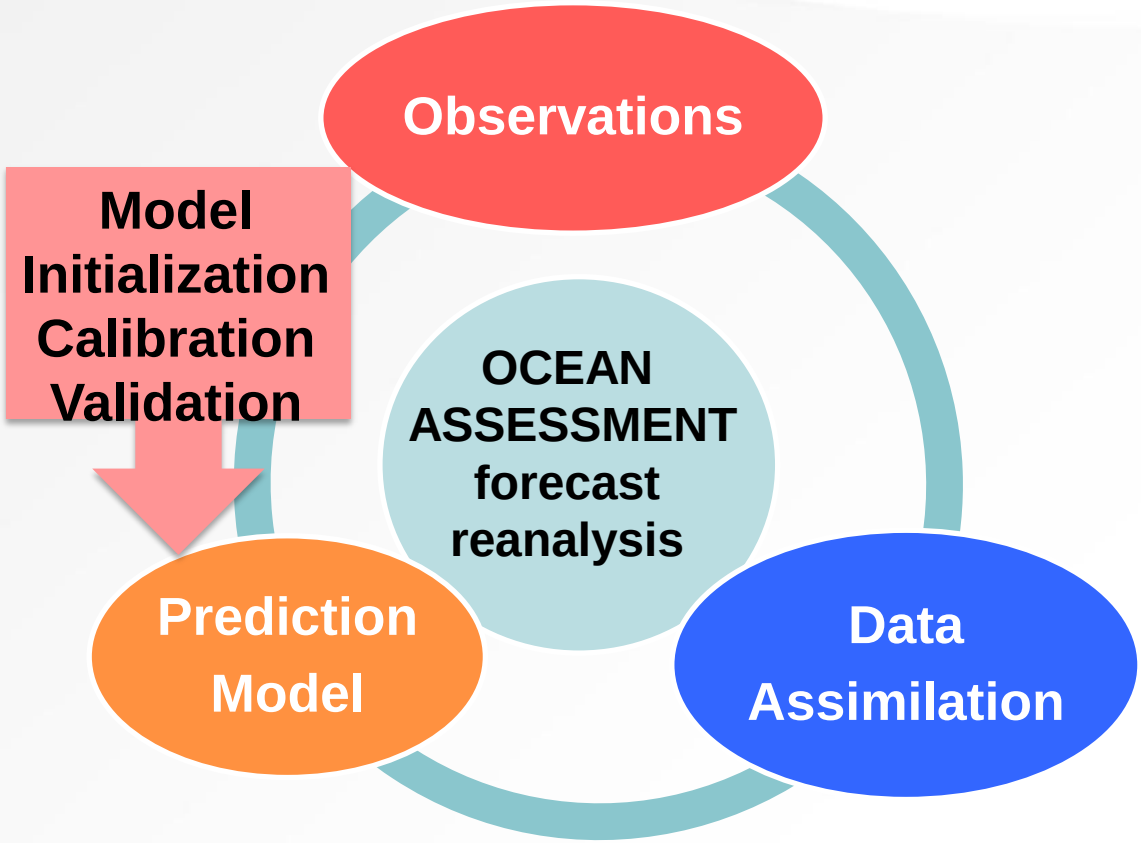
SDN-MyO COLLABORATION

Obj: to create TS historical data collections in synergy with MyOcean In-Situ Thematic Assemble Centre (INSTAC) to support and promote monitoring, modeling and downstream service development

SeaDataNet

MoU

myOcean



sdn-userdesk@seadatanet.org – www.seadatanet.org

PAN-EUROPEAN INFRASTRUCTURE
FOR OCEAN & MARINE DATA
MANAGEMENT

SeaDataNet

SDN-MyO COLLABORATION

SeaDataNet

MoU

myOcean

1st SDN-MyO Joint Meeting

In order to generate jointly Regional TS products for 20 years (1990-2010) reanalysis planned in MyO2 and for SDN2 aggregated products we had to **define** the **interactions between the two Projects:**

- common time schedule
- data flow
- information exchanges
- QC strategies
- interfaces

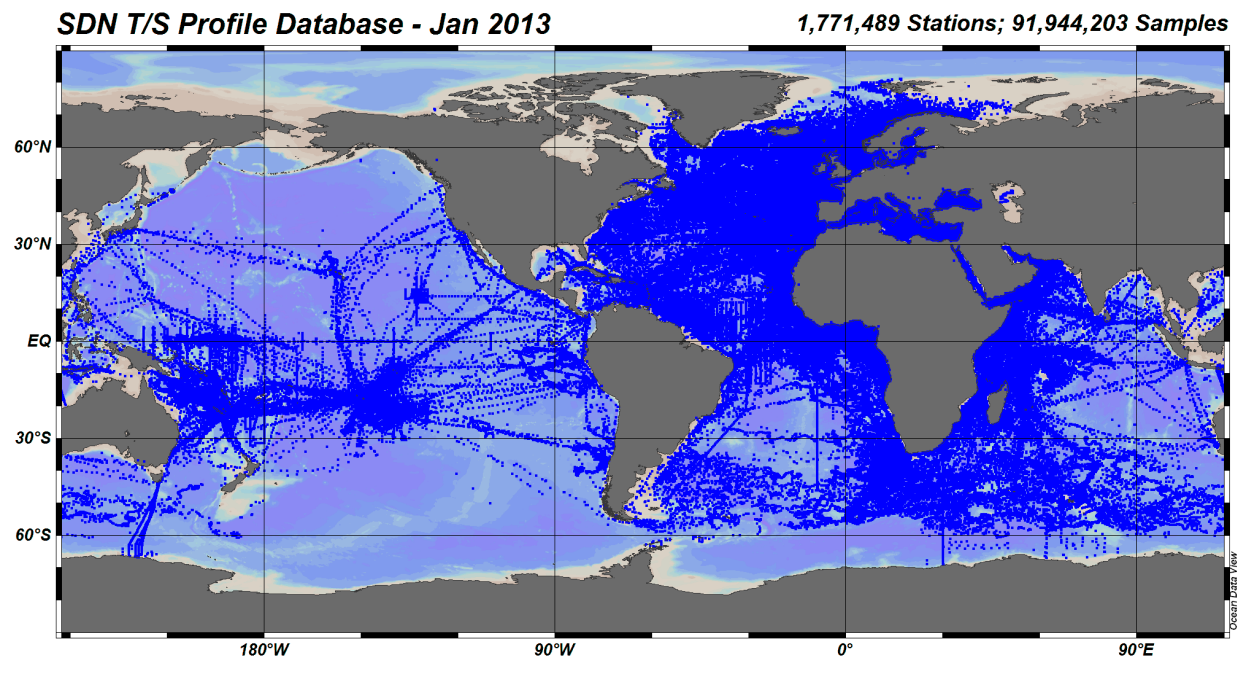
Aggregation Procedure

Data aggregation needed the **coordination and collaboration between different WPs**:

- **WP4** duplicate implementation plan (HCMR)
- **WP5** CDI robot harvesting (MARIS)
- **WP9** aggregation into a single TS Data Collection using SDN Importer of ODV 4.5.3 (AWI)
- **WP10** quality assessment by RCs in collaboration with MyO INSTAC
- **WP7** network of National Oceanographic Data Centres (NODC) → corrections of data anomalies

*S. Simoncelli, C. Coatanaon, Ö. Bäck, H. Sagen, S. Scory, D. Tezcan, D. Schaap, R. Schlitzer, S. Iona, M. Fichaut, M.Tonani. "Temperature and Salinity historical data collections for the European Merginal Seas: AGGREGATION AND QUALITY ASSESSMENT PROCEDURES". **IMDIS Conference 2013, Lucca, Italy***

~860.000 CDIs >2 Mio SDN data files in ODV format



~61000 potential duplicates where checked (6% real duplicates, 65% needing correction, 29% not duplicates)

>14000 files rejected because ODV was not standard or SDN standard

→ most of the data have been corrected

TIME SCHEDULE and ACTIONS

SeaDataNet

myOcean

SDN2		MyO2
Dupplicate implementation Plan	Oct2012	
Data Harvesting and Aggregation	Dec2012	
RCs received TS collections: basic QC	Jan2013	
	Feb2013 Release 1990-2010 aggregated data sets to MyO →	
RC Meeting	Apr2013 2° SDN-MyO Joint Meeting on QC	
WEBEX on AGGREGATION	May2013	← QC feedback to SDN RCs
Organization of QC feedback to NODCs	Jun2013	
QC Guidelines for NODCs	Jul2013	
NODCs QC actions dateline	15 Sept2013	

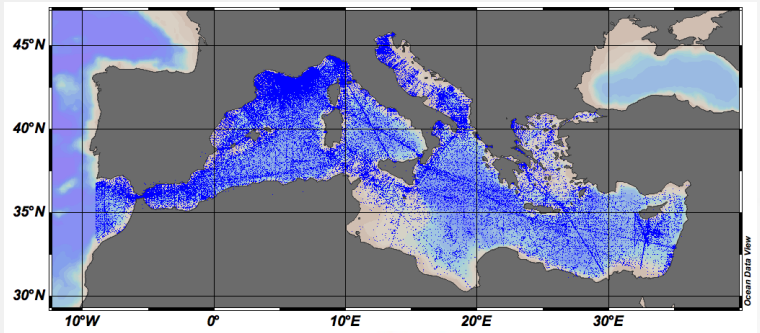
AGGREAGATED DATA SETS RELEASE
POSTPONED → Jan/Feb 2014

PAN-EUROPEAN INFRASTRUCTURE FOR OCEAN & MARINE DATA MANAGEMENT

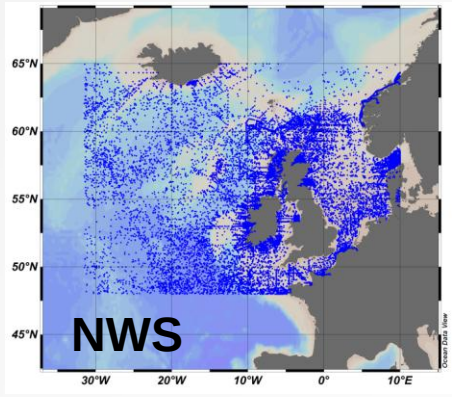
SeaDataNet

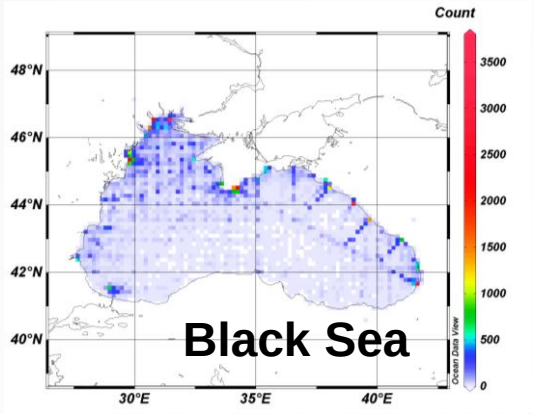
Aggregated Data Sets

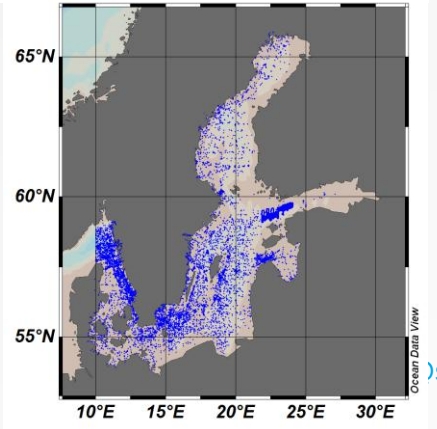
T&S historical data collections 1900-2010 were created for each European marginal sea to meet ***operational oceanography*** and ***climate change community*** requirements that need longer and longer time series of in situ observations to study long term ocean phenomena and their implications in the surrounding environment



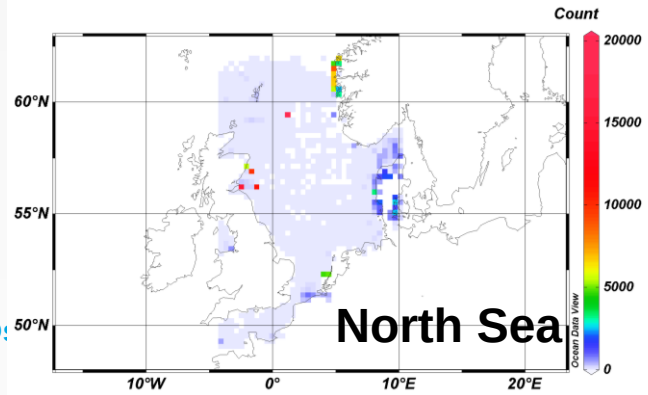
Mediterranean Sea

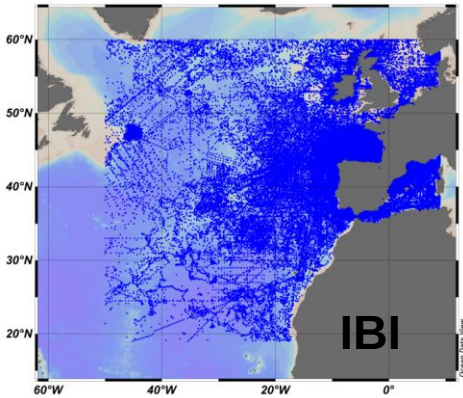






Baltic Sea





Second Year Activities

Jan2013: RCs received the TS collections → **guidelines for a first basic QC analysis in ODV and a common template for the QC report**

Feb2013: 1990-2012 sub sets of data have been extracted and released to MyO In-situ TAC in order to collaborate on the QC process

Mar2013: reports on the entire data collection have been prepared and presented to SDN StComm. StComm nominated a responsible person to coordinate the comunication between RCs-MyO INSTAC-NODC→ *Christine Coatanon (Ifremer)*

10Apr2013: RCs WEBEX Meeting on QUALITY ASSESSMENT

15Apr2013: 2nd SDN-MyO Joint Meeting on data quality assessment

May2013: MyO sent to SDN feedback on the quality of regional TS collections and anomalies files

Second Year Activities

Jul2013: analysis of MyO anomalies and subdivision of anomalies per CDI partner (*C. Coatanoan*)

27Jun2013: WEBEX on the Products (RCs, StComm members involver in the aggregation procedure) → Common decision to repeat the data aggregation exercise

Jul2013: list of anomalies sent to each CDI partner asking to take actions in order to correct possible errors

15Sept2013: CDI partners dateline to answer to anomalies → make corrections in the CDI

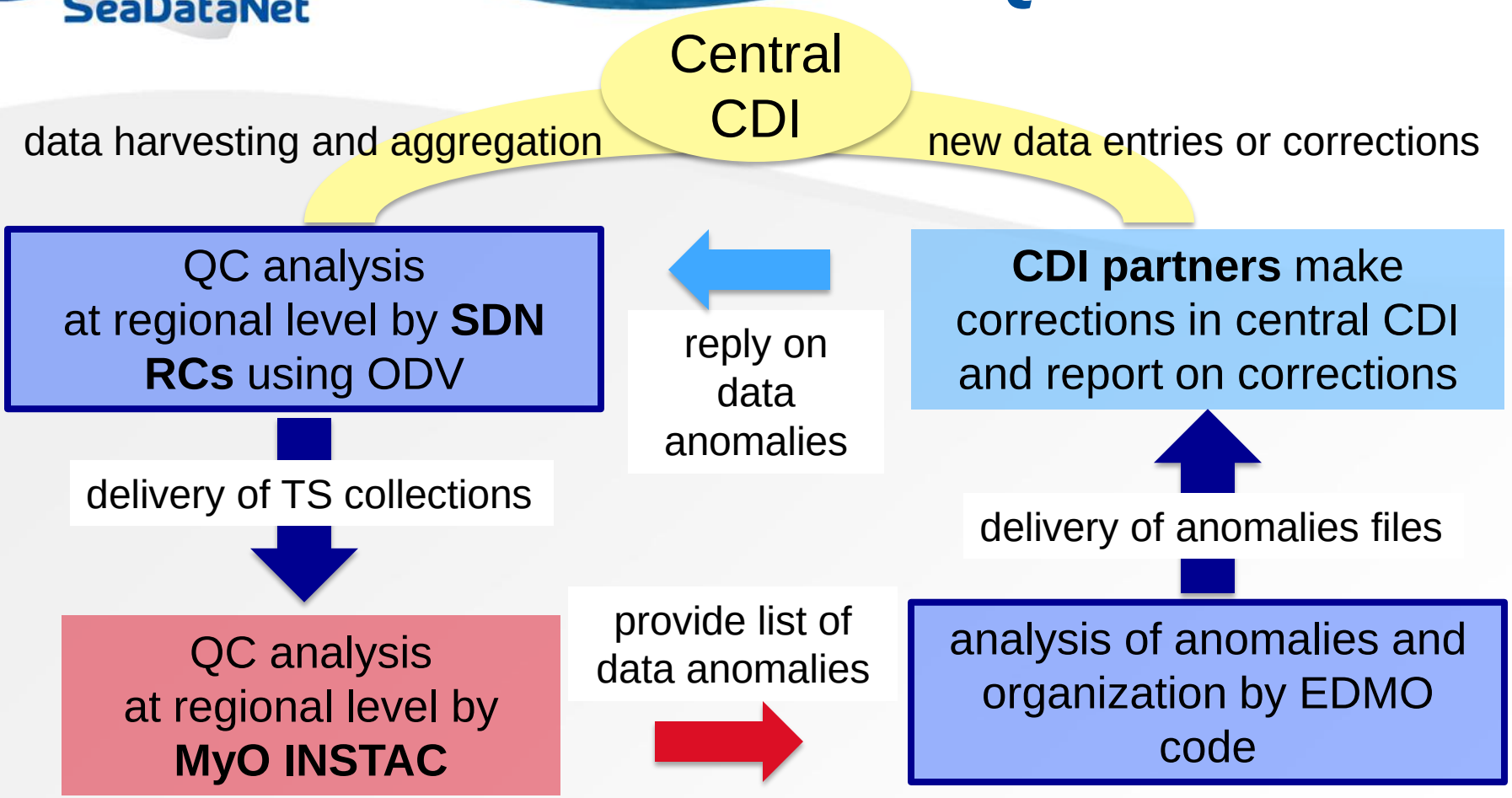
Nov2013: start of the new data aggregation exercise

→ **regional TS aggregated data sets release is postponed to Jan/Feb 2014**

→ **D10.2 postponed too (Draft is circulating between partners)**

SeaDataNet

QC Procedure



QC procedure has been designed to be iterative in order to facilitate the update and improvement of SDN database content

QC analysis

QC analysis considered all Qflags in order to identify anomalies and possible solutions and results were included in short reports

1. Data distribution and data density map
2. Histograms of annual and seasonal data distribution
3. TS scatter plots of the entire dataset→ highlighted the necessity of applying a gross range check
4. TS scatter plot after the **range check**
5. TS scatter plots of: Qflags=1(good),2(probably good) and Qflags=0(no check)
6. Statistics about Qflags
7. Visual control of scatter-plots to identify wrong profiles (outliers)
8. Visual check of spikes
9. Identification of stations falling on land, of wrong or missing data

Outliers have been saved in text files in order to report to both MyO and NODCs

QC Results

- Some data center probably inverted QF1 and QF0 → obs flagged as good (1,2) presented values out of range while most of the obs flagged 0 were reasonable → we asked for checking
- RCs decided to improve the strategy for future QC analysis through specific sub-regional QC (areas & depth) and stability check on density
- RCs will finalize with an harmonized approach the Reports on Quality Assessment

RCs do not modify data or QF but define procedures and priority actions to report on the quality to data providers in order to promote the update process and the progressive improvement of the infrastructure.

RCs were all active and collaborative in QC assessment and reporting activities. They participated to online meetings and to the Joint Meeting with MyOcean INSTAC.

MyO QC Results

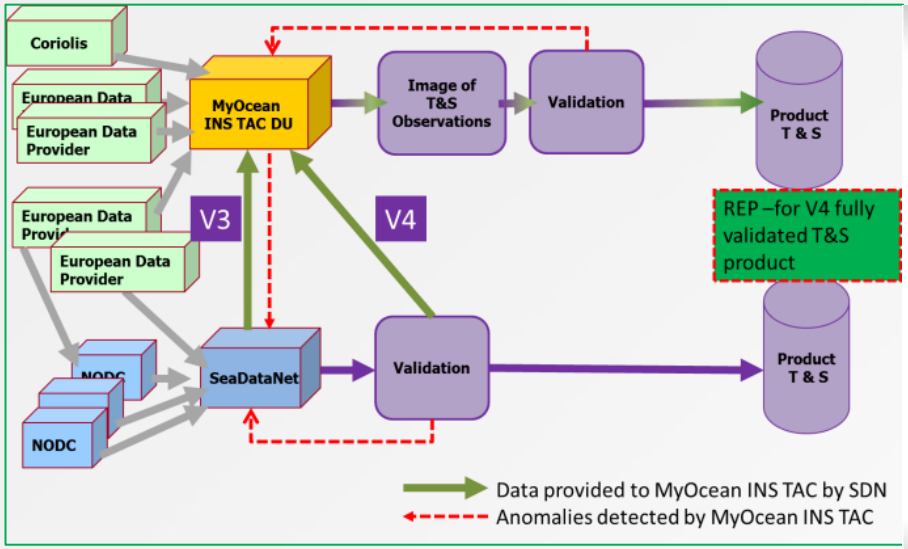
Main issue was that some important metadata have been lost in the aggregation process (i.e. Instrument Type for XBT).
INSTAC:

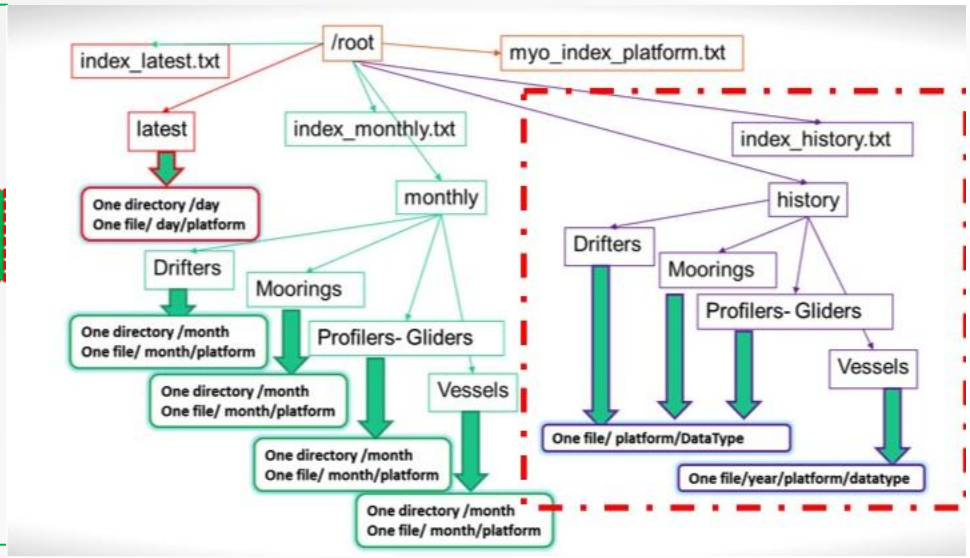
- decided to insert only new data and put aside the data that were detected as duplicates;
- detected some data with QF0 (no control) mixed with QF1 (good) either within a profile or in a series of profiles;
- asked to know whether CTDs are calibrated or not (criteria to replace RT data with DM data)

Globally the data were considered good but a visual QC was useful for checking suspicious data.

MyO Products

MyOcean V3 product → aggregation of the data from the ROOS providers and SDN removing duplicates and converting all data in the same format with the same QC flags





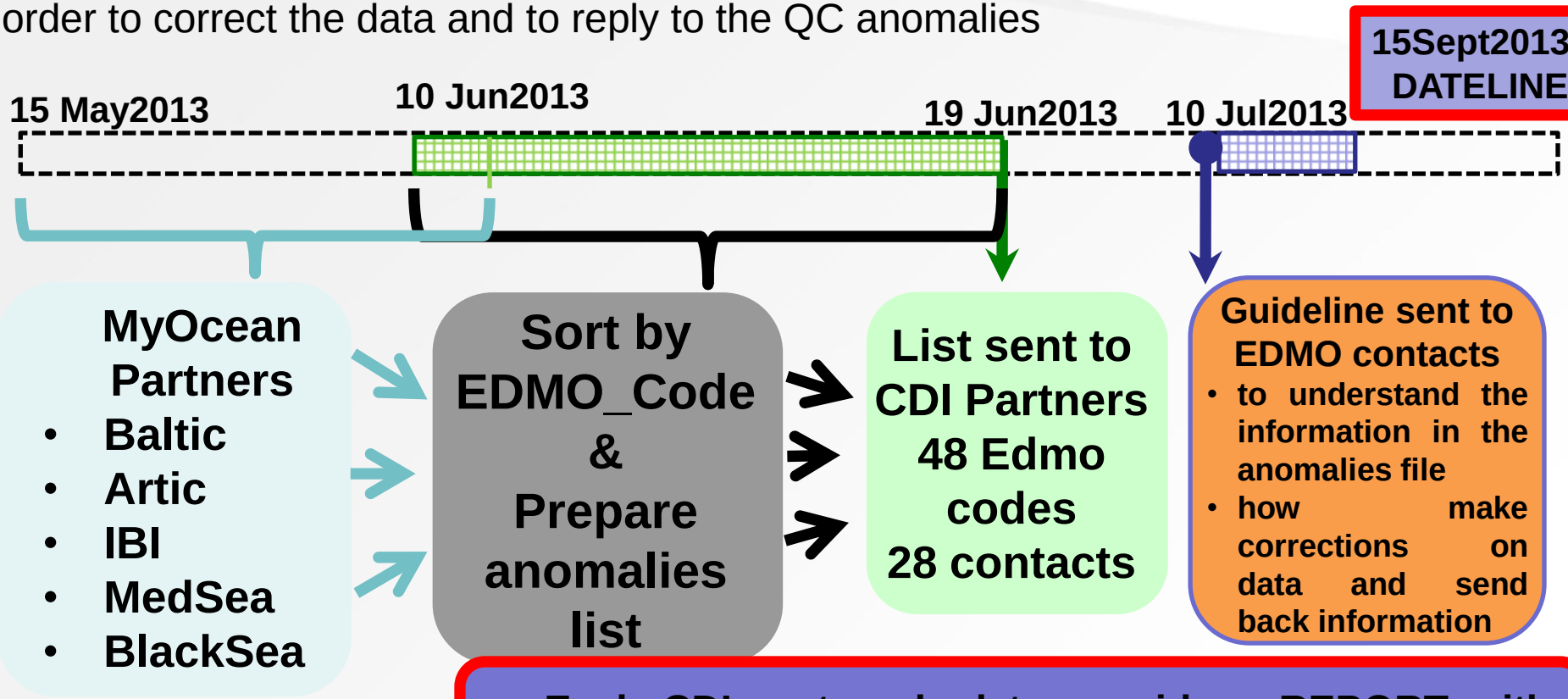
MyOcean V4 product → will include new and better quality SDN data

SDN-MyOcean Joint Product will be created as final result of this big collaboration effort

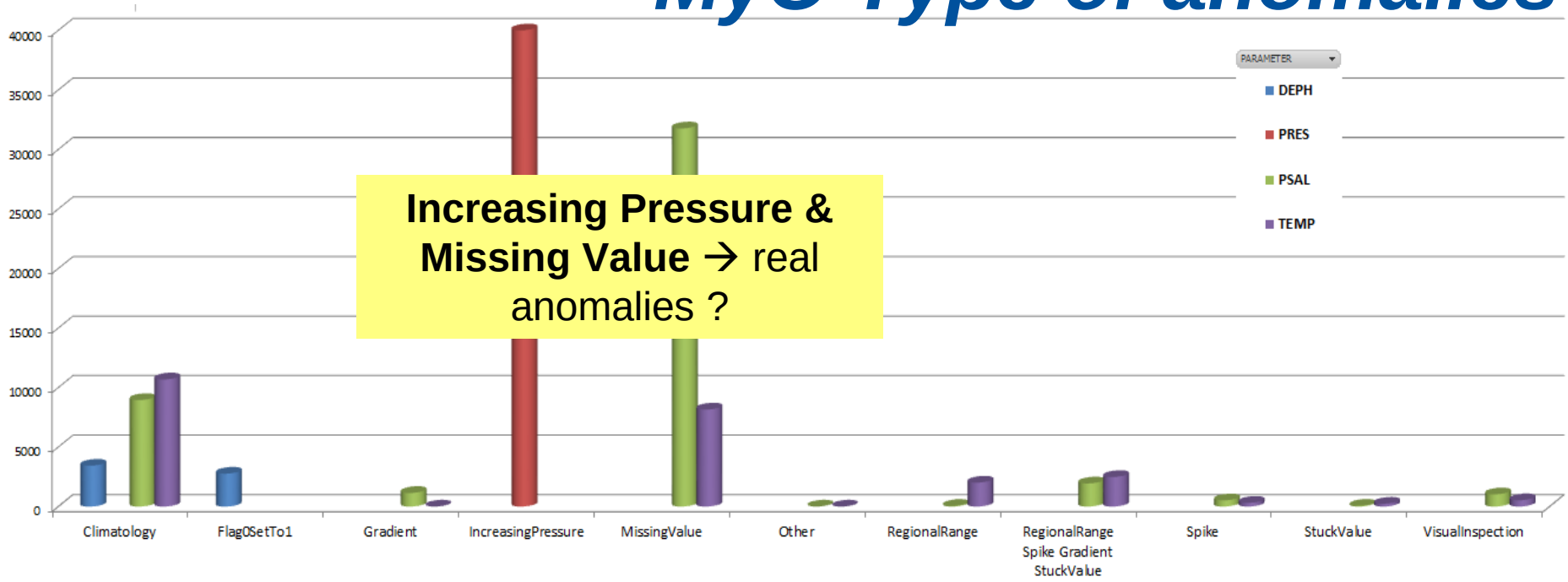
Exchanges SDN-MyO

Jun2013: Results from SDN and MyO QC analysis were collected, organized and sent to the NODCs

JulAug2013: NODCs were asked to check QC data anomalies, to take actions in order to correct the data and to reply to the QC anomalies



MyO Type of anomalies



Increasing Pressure & Missing Value → real anomalies ?

Climatology: out of the climatology envelope
Gradient: gradient anomaly
Increasing Pressure: non increasing depth
Missing Value: missing value with a Qflag≠9

Regional Range: value out of regional range
Spike: spike
Stuck Value: constant profile
Visual Inspection: finding an anomaly visually

Increasing pressure & Missing Value

"Missing Value" 0 to 9 came from a bug in ODV during export → corrected

"Pressure increasing" can have several origins:

- **Bottle platform** → no anomaly! Normal to have several samplings at the same depth
- **Duplicate CDIs problem:** in some cases data providers forwarded multiple updates of CDIs in a short time frame with overlaps→ not a true anomaly
- **True error** of non-increasing depth/pressure and then it should be flagged to 4

What is expected from each CDI partner ?

1) **List of anomalies** updated with NODC comments following this table

Column	Description	Comment
LOCAL_CDI_ID	*cdi_identifier*	Partner local CDI identifier, Information from CDI
EDMO_CODE	EDMO_CODE of the organization distributing the data	Information from CDI
PLATFORM_CODE=CRUISE	CDI cruise_name	
STATION_DATE_START	Date at which the station starts	
STATION_DATE_STOP	Date at which the station ends	
UPDATE_DATE	Date of the control done by MyOcean partners	
PARAMETER	PARAMETER exported from ODV (TEMP, PSAL, DEPH or DEPTH [sometimes PRES when MyOcean partners have changed name])	
QC_ACTION	As described in Introduction, to define the type of anomalies (spike, gradient, missing value, etc)	
OLD_QC	QC from original dataset	
NEW_QC	QC suggested by MyOcean (see Annex I)	
VERTICAL_REFERENCE_START	Level at which starts the anomaly in the profile	
VERTICAL_REFERENCE_STOP	Level at which stops the anomaly in the profile	
AGREE WITH THE SUGGESTED CORRECTION (YES/NO)	Fill with Yes/No if you agree/disagree with the corrections suggested by MyOcean	
NODC COMMENT	Column to be added to your file in order to put some information about your our opinion about the suggested correction (agreement, disagreement, explanation if necessary)	
DETAILS	Column to be added to your file in order to put more information about suggested correction	

2) **List of updated CDI**

LOCAL_CDI_ID	EDMO_CODE	PLATFORM_CODE=CRUISE
FI35199101301_00050_H10	486	PRIMO-0 21/03
FI35199443005_25900_H10	486	MBP-FRONT 1994
FI35199502002_00870_H10	486	EUROMARGE
FI35199706005_0K010_H10	486	PELMED 97
FI35199845001_00260_H10	486	BIODYPAR 1

3) **report** with:

- list of errors and number
- details on why corrections have not been applied, etc……

QC_Action	Number of anomalies detected by MyOcean	Number of true anomalies	%
Climatology	20	0	0.0
Gradient	328	1	0.3
IncreasingPressure	544	25	4.6
RegionalRange	42	0	0.0
Spike	148	42	18.3
StuckValue	28	0	0.0
VisualInspection	1	1	100.0
Total	**1111**	**69**	**6.2**

Conclusions

- **Data Aggregation Procedure** was an extensive and fruitful exercise to manage more than 1 Million data and involving many WPs, people and institutions;
- **Data Aggregation** was huge distributed effort involving 62 data centers and more than 300 data originators. It permitted to identify and correct lots of data.
- **Quality Assessment Procedure** took place in coordination between SDN RCs - MyO INSTAC - SDN CDI partners
- **Aggregation and Quality assessment** allowed to **ameliorate and refine each technical phase** but mainly **to highly improve the quality of SDN infrastructure content**
- Collaboration between SDN and MyO was crucial during the data quality assessment and allowed **to identify and correct lots of data anomalies**
- WP10 promoted collaborations and communication between partners, WPs (WebEX conf) and projects (2 Joint Meetings)

Conclusions

Since the quality of historical data collections would be highly improved by an update before the official release the StComm and RCs agreed to repeat aggregation procedure and QC assessment

A **new aggregation exercise** allows to

1. deliver the best aggregated data sets
2. refine QC procedures
3. retrieve also **restricted data** to be used for the statistical products computation (Due M36)
4. increase the quality of SDN statistical products (climatologies, maps, profiles) based on data collections
5. Reinforce and renew collaboration with MyO: improve MyO RT QC procedures with SDN statistical products

→ D10.2 has been postponed to February 2014 with the official release of V1 aggregated data sets

CDI

TS aggregated dataset
free-acces data + under SDN licence + **restricted data (new!)**

A **new data aggregation exercise** allows to

1. deliver the best aggregated data sets which mirror the true infrastructure content
2. retrieve also **restricted data** that have to be used for the internal products computation
3. SDN products based on data collections will be higher quality

REGIONAL TS COLLECTIONS

Arctic **Noth Atlantic** **North Sea** **Baltic Sea** **Med Sea** **Black Sea**

1900-2012
COMPLETE DATA COLLECTIONS
(containing restricted data)
for internal use only
→ SDN STAT PRODUCTS

1. 1900-2012
DATA COLLECTIONS
(unrestricted data only)
for external SDN release
2. 1990-2012 SUB-SETS MyOcean

Statistical Products

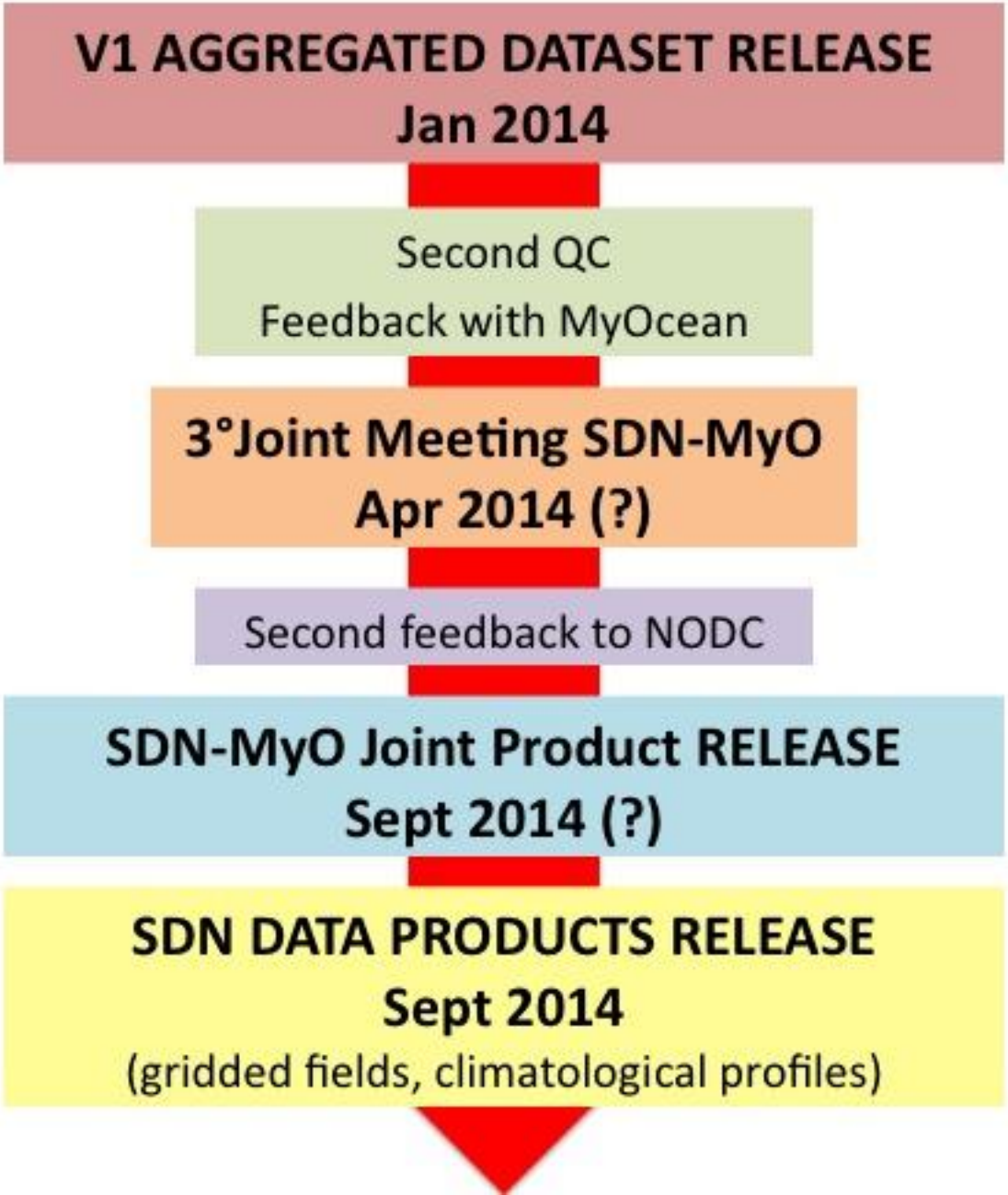
1. show and synthesize the SDN database content
2. used to implement new common QC analysis in synergy with MyO
3. serve external users for both data analysis and modeling purposes

T&S monthly climatologies and corresponding error field for all regional seas on IODE standard levels

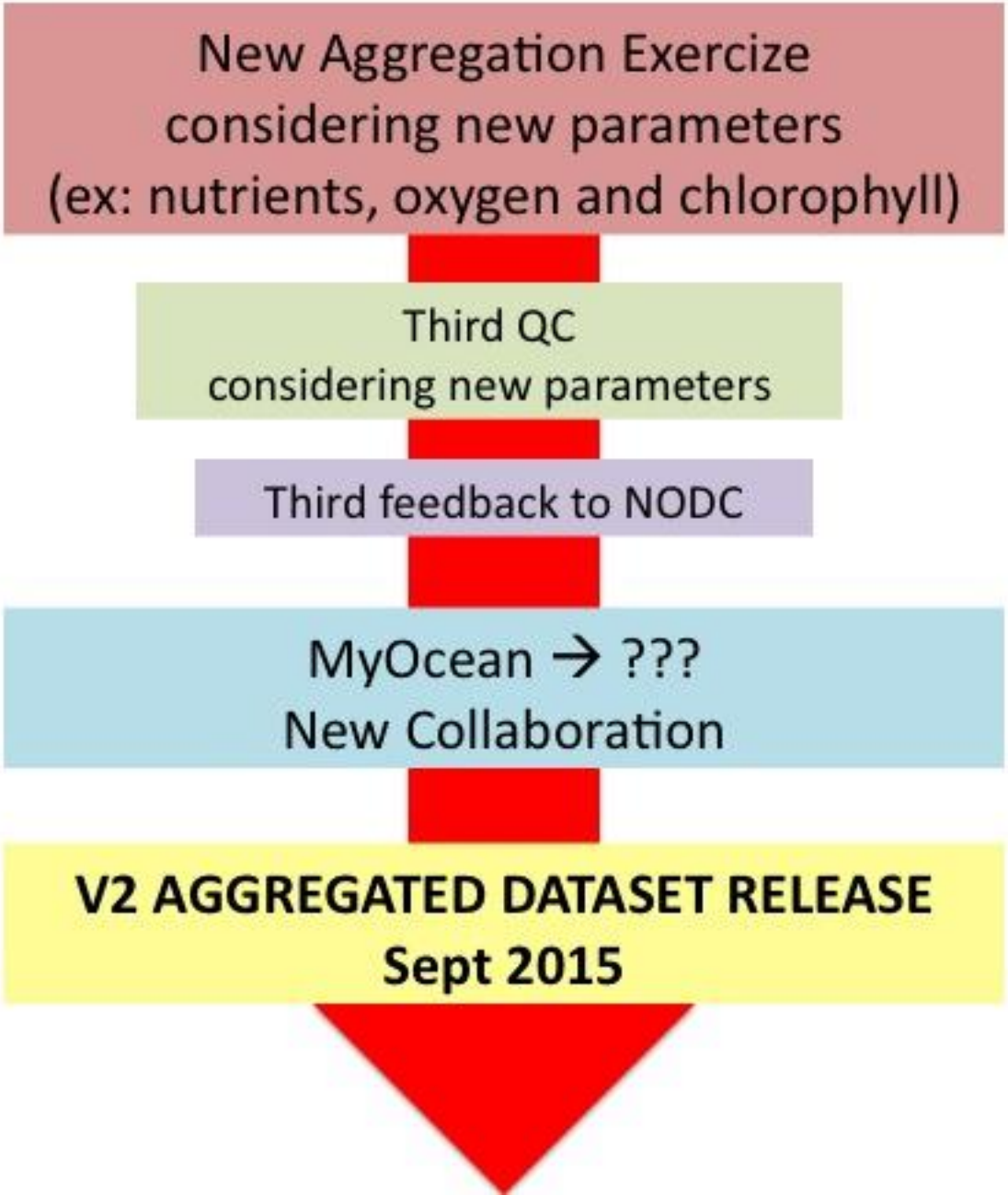
TS Collections for climatology computation will contain restricted data and will be integrated with external data sources.

Other parameters will be taken into consideration later in the project (nutrients, oxygen and chlorophyll)

WP10 Work Plan 2014



WP10 Work Plan 2015



Thank You