S.Simoncelli, M.Tonani
D. Shaap, R. Shlitzer, S. Iona, M. Fichaut, C. Coatanon, O. Back, S. Scory, H. Saegen, D. Tezcan

SeaDataNet

PAN-EUROPEAN INFRASTRUCTURE
FOR OCEAN & MARINE DATA
MANAGEMENT

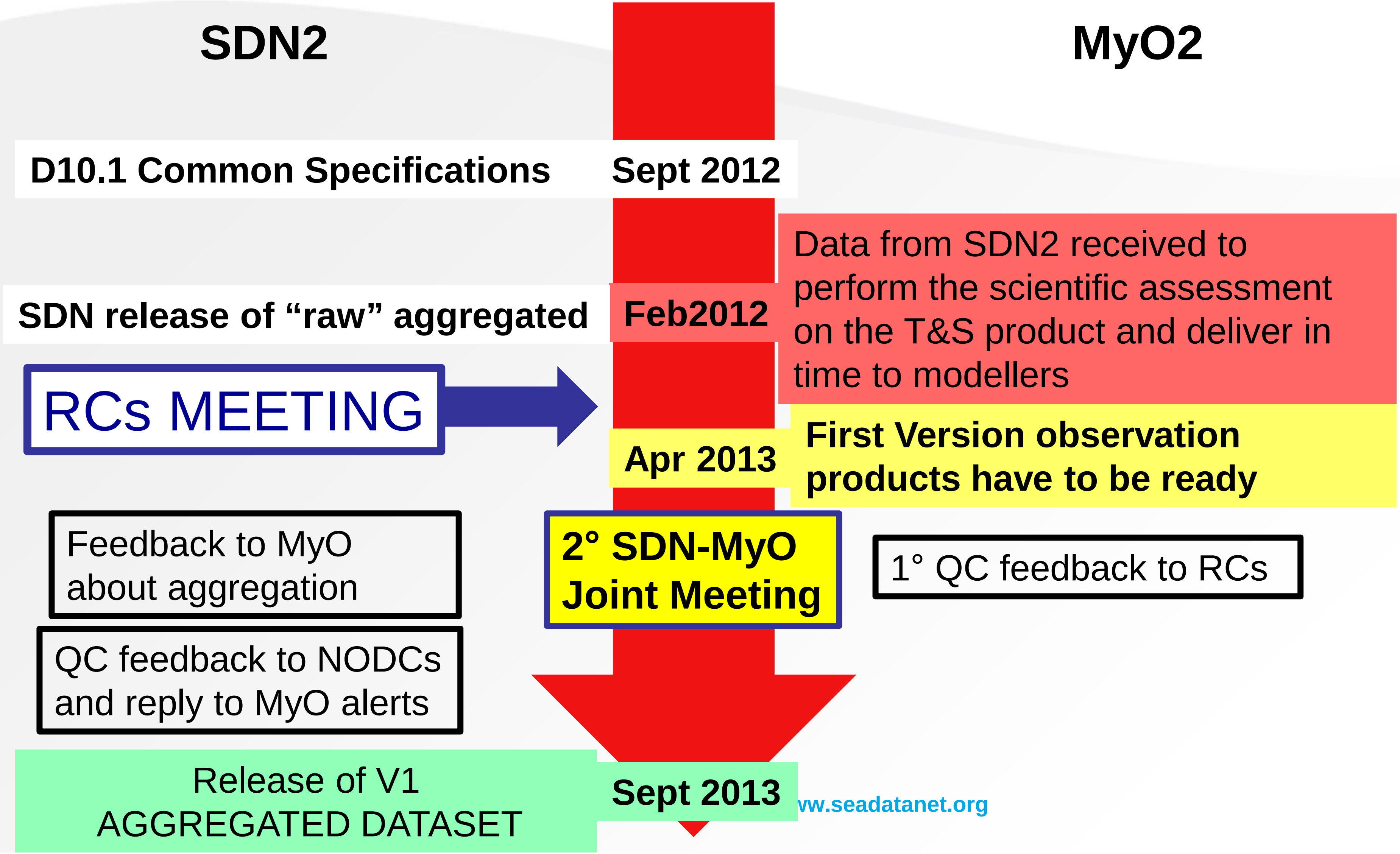
SDN feedback on dataset aggregation: What worked well What have been the difficulties

2nd SDN – MyO Joint Meeting, Cork April 15th, 2013

OUTLINE

- Common SDN-MyO Time Schedule Review
- Harvesting of T&S files by CDI Robot (*D. Shaap-MARIS*)
- Building SDN Aggregated Datasets (R. Schlitzer-AWI)
- Duplicates Issue (*S. Iona-HCMR*)
- Correction of ODV files (*M. Fichaut*)
- Regional Coordinators preliminary QC analysis
- Coming next
- Conclusions

TIME SCHEDULE REVIEW



SDN2

MyO2

D10.1 Common Specifications — Sept 2012

Data from SDN2 received to perform the scientific assessment on the T&S product and deliver in time to modellers

SDN release of “raw” aggregated — Feb2012

RCs MEETING

Apr 2013 — First Version observation products have to be ready

Feedback to MyO about aggregation

2° SDN-MyO Joint Meeting

1° QC feedback to RCs

QC feedback to NODCs and reply to MyO alerts

Release of V1 AGGREGATED DATASET — Sept 2013

Harvesting of T&S files by CDI Robot *(D. Shaap)*

- MARIS developed a **Robot user** that uses the CDI Data Discovery and Access Service **to query, shop and retrieve data sets from the distributed data centres in an automatic way**
- Query for the joint product → search for all data sets with T&S and for which the access restriction is Unrestricted or SeaDataNet License => ca 860.000 CDIs
- then the Robot was triggered to start harvesting the related ODV files from the distributed data centres through the general CDI shopping mechanism (RSM – DM)
- This was also used to test and tune the performance of the RSM – DM process to find the optimum data requests

Harvesting of T&S files by CDI Robot *(D. Shaap)*

- All data requests are administered in the RSM
- For some data centres the Robot had to download more than 100.000 files, but it is not possible to process such a large request in one go by a DM, because of memory problems and internal 10min clock cycle
- At the start RSM was set to slice large requests to 5000 data sets per cycle of 10 minutes.
- Processing 100.000 datasets from 1 data centre would then take theoretically 20 times 10 minutes
- However the 5000 datasets slice caused memory problems at specific data centres → tuning took place and finally the slicing factor was set at 500 data sets per cycle of 10 minutes which can be handled by all connected data centres.

Harvesting of T&S files by CDI Robot *(D. Shaap)*

- Retrieving 100.000 datasets from 1 data centre can thus be done by the Robot through RSM–DM in parts of 500 per 10 minutes → implicating a total retrieval period of 100.000 / 500 = 200 * 10 minutes.
- RSM is fault proof => it keeps track of all data requests and repeats data requests in case of disturbances at DM level (the DM can be considered as slave with little intelligence, while RSM is master)
- Robot harvesting and tuning of the shopping system → mid Dec2012 taking into account also the **duplicates issue**
- A DVD was prepared for AWI (*R. Schlitzer*) with all retrieved ODV files in a storage structure with the full CDI metadata as CSV file and including a path per CDI to the related ODV data set on the DVD
- DVD delivered to AWI mid Jan2013. The ODV files contained in most cases not only T&S but also additional observations

Building SDN Aggregated Datasets (R.Schlitzer)

- >2 Mio SDN data files in ODV format
- metadata file containing CDI information for all

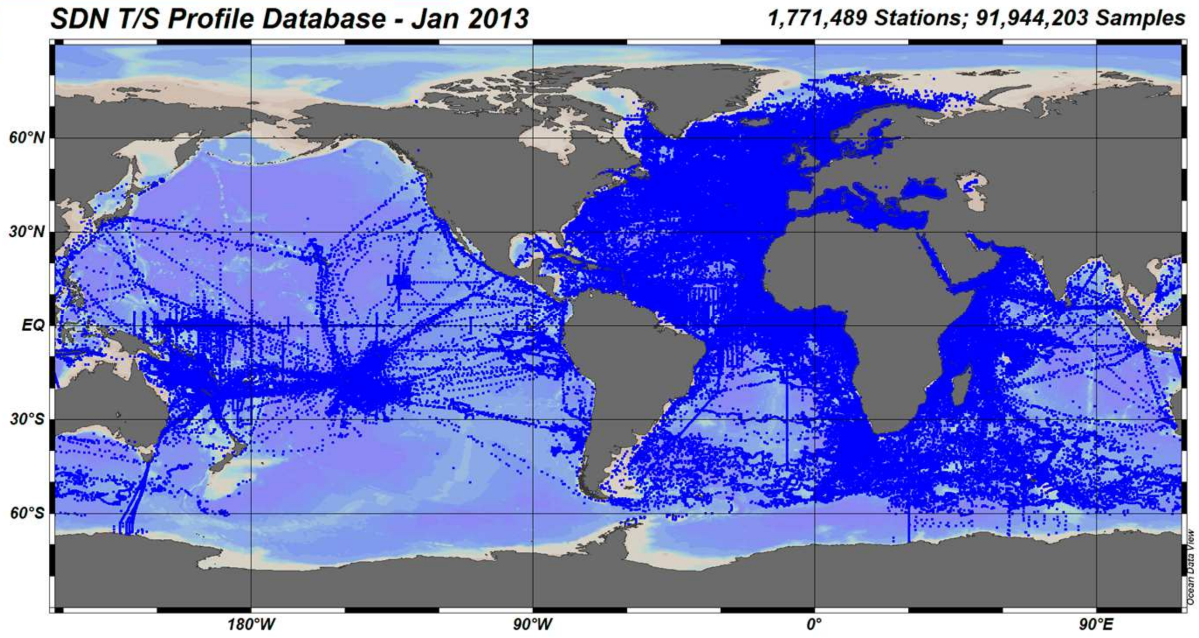
Aggregation of all Data Files into Single TS Data Collection

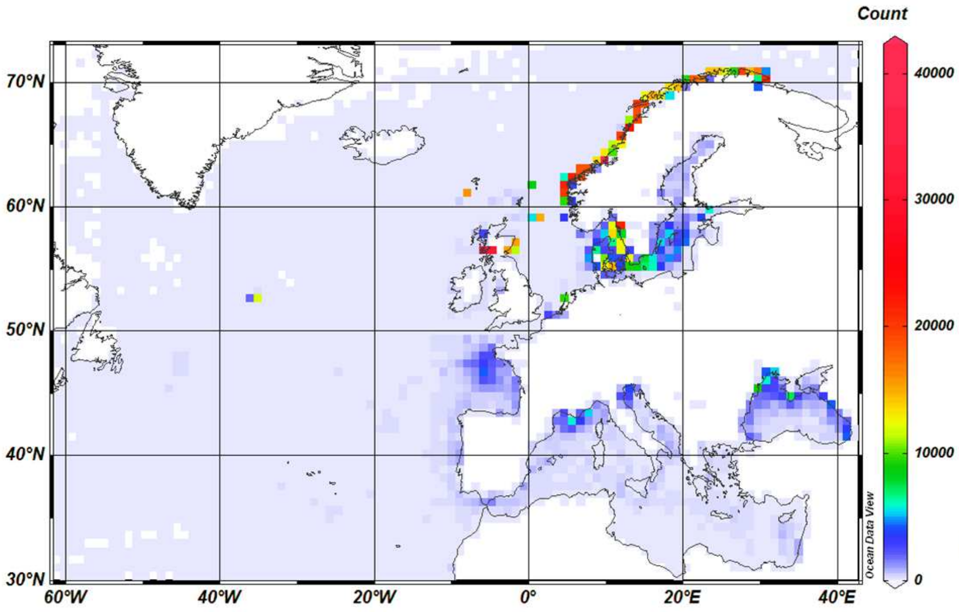
- Using SDN Importer of ODV 4.5.3
- Done in 9 pieces of about 250,000 files each, then combined
- Aggregation of the many original temperature and salinity variables into single T and S variables using „Aggregated Derived Variables“
- Analysis logs of problem files sent to coordinator/data centers for fixing
- Creation of regional and 1990-2012 subsets and distribution to SDN regional groups

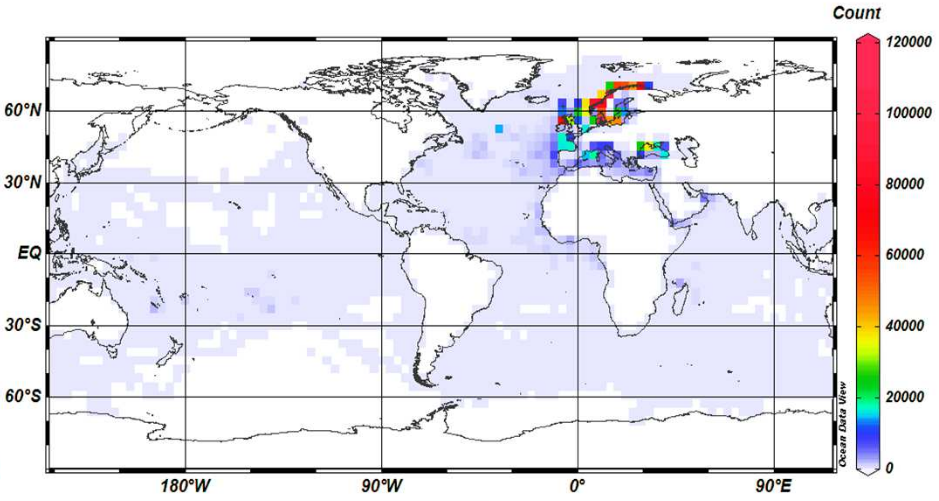
PAN-EUROPEAN INFRASTRUCTURE
FOR OCEAN & MARINE DATA
MANAGEMENT

SeaDataNet

(R.Schlitzer)


SDN T/S Profile Database - Jan 2013
1,771,489 Stations; 91,944,203 Samples
60°N
30°N
EQ
30°S
60°S
180°W
90°W
0°
90°E
Ocean Data View





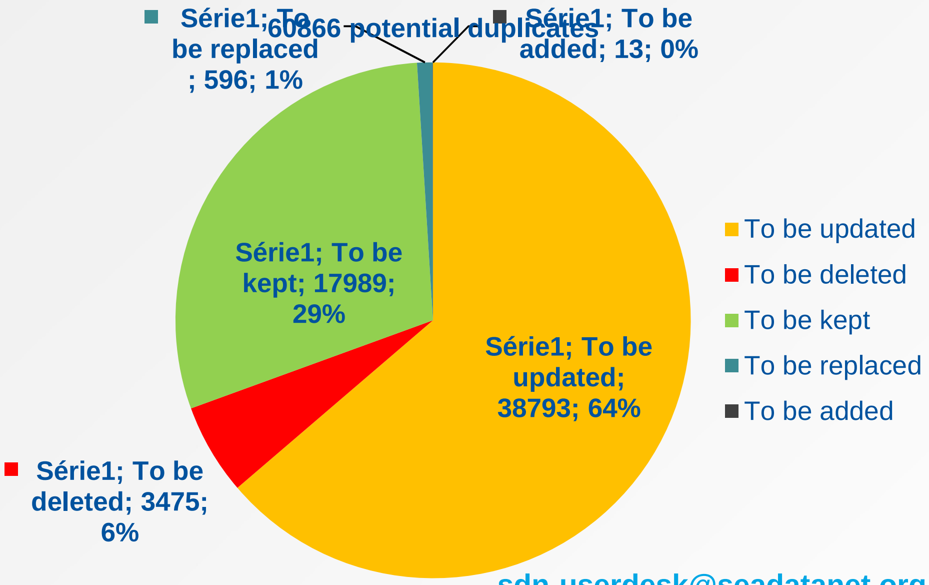
Duplicates Implementation Plan

- Based on the duplicates checks conducted by ODV for the 6 SDN data coverage regions, an ***implementation plan*** was prepared and sent to all SDN partners (on early Oct2012), asking for:
  - ✓ identification of duplicates
  - ✓ cleaning of their data sets (delete, update, replace, etc)
  - ✓ detailed explanations of their actions
- After evaluation of the modifications of each partner, the CDI central catalogue (as well as the local archives) was updated accordingly
- There are some missing cases (partners who still process their data sets)

Results

Number of duplicates and actions undertaken:

21 Partners	Potential duplicates	To be updated	To be deleted	To be kept	To be replaced	To be added
Total	60866	38793	3475	17989	596	13



Conclusions:

✓The majority of potential duplicates (71%) were real duplicates (6%) or needed correction (65%).

✓Only 29% were not duplicates and remained as they were.

Main reasons & explanations provided by data providers

- **Deleted CDIs:** one data distributor has submitted parts of the same data set (74%)
- **Updated CDIs:** data sets included unknown, wrong, missing information or partners have submitted false data sets (98.5%)
- **Replaced CDIs:** submission of data sets with unknown, wrong, missing Information or the submission of false data sets(95%)
- **Kept CDIs:** the majority of CDIs that were found as potential duplicates were in fact replicates because of unknown time, time-space differences less the threshold values, different measurement methods (83.5%)

Next Actions for duplicates check

- ***Guidelines*** will be sent to partners (new and old ones) to avoid similar cases in the future.
- A ***white list*** of the cleaned and checked CDIs has been prepared
- ***New entries*** in the CDI central catalogue ***will be checked*** against this list to avoid future duplicates in the future

WP 4-5 – Corrections of ODV files (M.Fichaut)

- During the preparation of the aggregated dataset for MyOcean, more than 14 000 files were rejected because ODV was not Standard or not SDN standard
- ODV files had to be corrected
- List of errors sent to 33 data centres, among them 5 are not SDN partners (mid Feb2013)

Code	Distributed by	Name	Country	SDN partner	Nb files	Comments	Status	Test	Test comment
EDMO_43	BODC	BODC	UK	Yes	17	Missing variables in th SDN header Bot.Depth to be replaced by Bot. Depth	mail 11/02/2013		
EDMO_100	BSH	Baltic Sea Research Institute Warnemuende (IOW)	Germany	No	433	Missing variables in th SDN header	Corrected 13/02/2013	OK	
EDMO_108	OGS	CNR, Istituto di Scienze Marine (Sezione di Venezia - ex IBM)	Italy	No	1745	Missing variables in th SDN header	Corrected 14/02/2013	OK	
EDMO_120	OGS	OGS , Department of Oceanography	Italy	Yes	535	Missing variables in th SDN header	Corrected 14/02/2013	OK	
EDMO_127	OGS	CNR, Istituto di Scienze Marine (Sezione di Trieste)	Italy	No	1494	Missing variables in th SDN header	Corrected 14/02/2013	OK	
EDMO_134	ENEA	CNR, Institute of Marine Science U.O.S. of Pozzuolo di Lerici (SP)	Italy	No	157	Missing variables in th SDN header	Corrected 22/02/2013		
EDMO_136	ENEA	ENEA Centro Ricerche Ambiente Marino - La Spezia	Italy	Yes	273	Missing Bottom Depth column? Missing '//SDN_parameter_mapping' line Wrong postion of LOCAL_CDI_ID and EDMO_code comlun	Corrected 22/02/2013		
EDMO_144	OGS	Institute of Marine Science (ISMAR) - Ancona	Italy	No	70	Missing variables in th SDN header	Corrected 14/02/2013	OK	
EDMO_192	NIMH-BAS	Laboratory of Marine Ecology-Central Laboratory of General Ecology	Bulgaria	No	1	1 file: extra empty column?	mail 12/02/2013		
EDMO_237	OGS	Stazione Zoologica Anton Dohrn of Naples	Italy	No	251	Missing variables in th SDN header	Corrected 14/02/2013	OK	
EDMO_353	IEO	IEO	Spain	Yes	1	Duplicate PSAL parameter?	Corrected 11/02/2013	OK	
EDMO_396	MI	Marine Institute	Ireland	Yes	3	Pb with the header line	Corrected 12/02/2013	OK	
EDMO_486	IFREMER	IFREMER	France	Yes	12	Missing variables in the SDN header	Corrected 11/02/2013	OK	
EDMO_697	NIMRD	National Institute for Marine Research and Development "Grigore Antipa"	Romania	Yes	4	Missing variable in th SDN header	Corrected 12/02/2013	OK	
EDMO_698	LHEI	Latvian Institute of Aquatic Ecology	Latvia	Yes	36	Extra empty column?	mail 11/02/2013	OK	
EDMO_727	MHI	Marine Hydrophysical Institute	Ukraine	Yes	2	Pbs in the header	Corrected 13/02/2013	KO	Flag 0?
EDMO_732	BSTU	Karadeniz Technical University, Faculty of Marine Sciences	Turkey	No	29	Missing variables in th SDN header Extra empty column? EDMO_Code to be replaced by EDMO_code	Corrected 15/02/2013	KO	Still the same errors
EDMO_733	SNUFF	Sinop University, Fisheries Faculty	Turkey	No	126	Pb in the header line, one extra empty column?	mail 12/02/2013		
EDMO_840	IBSS	Institute of Biology of the Southern Seas, NAS of Ukraine	Ukraine	Yes	642	Inversion latitude - longitude?	Corrected 28/02/2013		
EDMO_989	BSH	Federal Research Centre for Fisheries (Cuxhaven)	Germany	No	21	Missing variables in th SDN header	Corrected 13/02/2013	OK	
EDMO_990	BSH	Federal Research Centre for Fisheries (Hamburg)	Germany	No	367	Missing variables in th SDN header	Corrected 13/02/2013	OK	
EDMO_991	BSH	Federal Research Centre for Fisheries Institute for Baltic Sea Fishery	Germany	No	126	Missing varaibles in the SDN header	Corrected 13/02/2013	OK	
EDMO_993	BSH	State Agency for Environment, Nature and Geology, Mecklenburg-Vorpommern	Germany	No	827	Missing variables in th SDN header	Corrected 13/02/2013	OK	
EDMO_1169	ONU	Odessa National I.I.Mechnikov University	Ukraine	No	25	Missing variables in th SDN header	mail 12/02/2013		
EDMO_1181	BSH	State Agency for Nature and Environment of Schleswig Holstein (LANU)	Germany	No	838	Missing variables in th SDN header	Corrected 21/03/2013		
EDMO_1265	TSU-DNA	Scientific - Research Firm "GAMMA"	Georgia	No	73	Missing variables in th SDN header	Corrected 21/02/2013		
EDMO_1327	BSH	Lower Saxony Water Management, Coastal Defense and Nature Conservation Agency	Germany	No	151	Missing variables in th SDN header	Corrected 13/02/2013	OK	
EDMO_1575	BSH	Federal Research Institute for Rural Areas, Forestry and Fisheries (VTI)	Germany	No	73	Missing variables in th SDN header	Corrected 13/02/2013	OK	
EDMO_1578	MUMM	MUMM, Belgian Marine Data Centre	Belgium	Yes	4581	Missing variables in th SDN header Duplicates variables in the column header	Corrected 25/03/2013		
EDMO_1850	BSH	Federal Maritime and Hydrographic Agency	Germany	No	983	Missing variables in th SDN header	Corrected 13/02/2013	OK	
EDMO_2121	TSU-DNA	Georgian Institute of Hydrometeorology of Georgian Technical University	Georgia	No	75	Missing variables in th SDN header	Corrected 21/02/2013		
EDMO_2122	TSU-DNA	Georgian Institute of Water Management of Georgian Technical University	Georgia	No	48	Missing variables in th SDN header	Corrected 21/02/2013		
EDMO_2537	BSH	State Office for Agriculture, Environment and Rural Areas of Schleswig Holstein (LLUR)	Germany	No	209	Missing variables in th SDN header	Corrected 13/02/2013	OK	
				Total	14228				

WP10: Export of Temperature and Salinity

- Data 1900-2012 have been extracted
- Dataset 1990-2012 have been released to MyO In-situ TAC
- Guidelines for a first basic QC (ODV) have been given
- A template for the report has been given
- Reports on the 1900-2012 dataset and 1990-2012 subset for MyO have been prepared and presenteded to the StComm

Basic QC STEPS

- Station Selection Criteria: 1/Jan/1990-31/Dec/2012
- Polygon Selection to avoid some areas
- Data distribution and data density map
- Histograms with annual and seasonal data distribution
- TS scatter plots of the entire dataset highlighted the necessity of applying a gross range check
- Scatter plot of T and S after the range check
- Scatter plot obs with QC flags 1 (good), 2 (probably good): obs flagged as good present values out of range!!!!
- Scatter plot obs with QC flags 0 (no quality check): there are many observations that did not pass through any QC procedure!!!!
- Outliers with respect to the defined ranges have been saved in text files in order to report to both MyO and the NODCs

QC Outcome

- T-S datasets require QC analysis regardless their QC flag to identify anomalies and possible solutions
- Statistics about QC flags
- Harmonize the first QC reports
- TS scatter plots of: 1) entire data set (before and after range check) ; 2) QC=1,2; 3) QC=0
- Visual control of scatter-plots to identify wrong profiles and outliers and visible spikes
- Identifying and marking stations falling on land
- Identify wrong or missing data

StComm and RCs Meeting Outcome

RCs will:

- not modify the data or the QC flags but define procedures to report to data providers in order to facilitate the update procedure and to progressively improve the quality of the infrastructure
- identify priority actions to be taken from the NODCs
- have a responsible person to coordinate the comunication between NODC-RC-MyO INSTAC → *Christine Coatanon* (Ifremer) with the help of *M. Fichaut* and *S. Iona (HCMR)*
- have a common strategy for future QC analysis: sub-regional QC (per areas & per depth), stability check on density
- Identify data providers having most problematic data

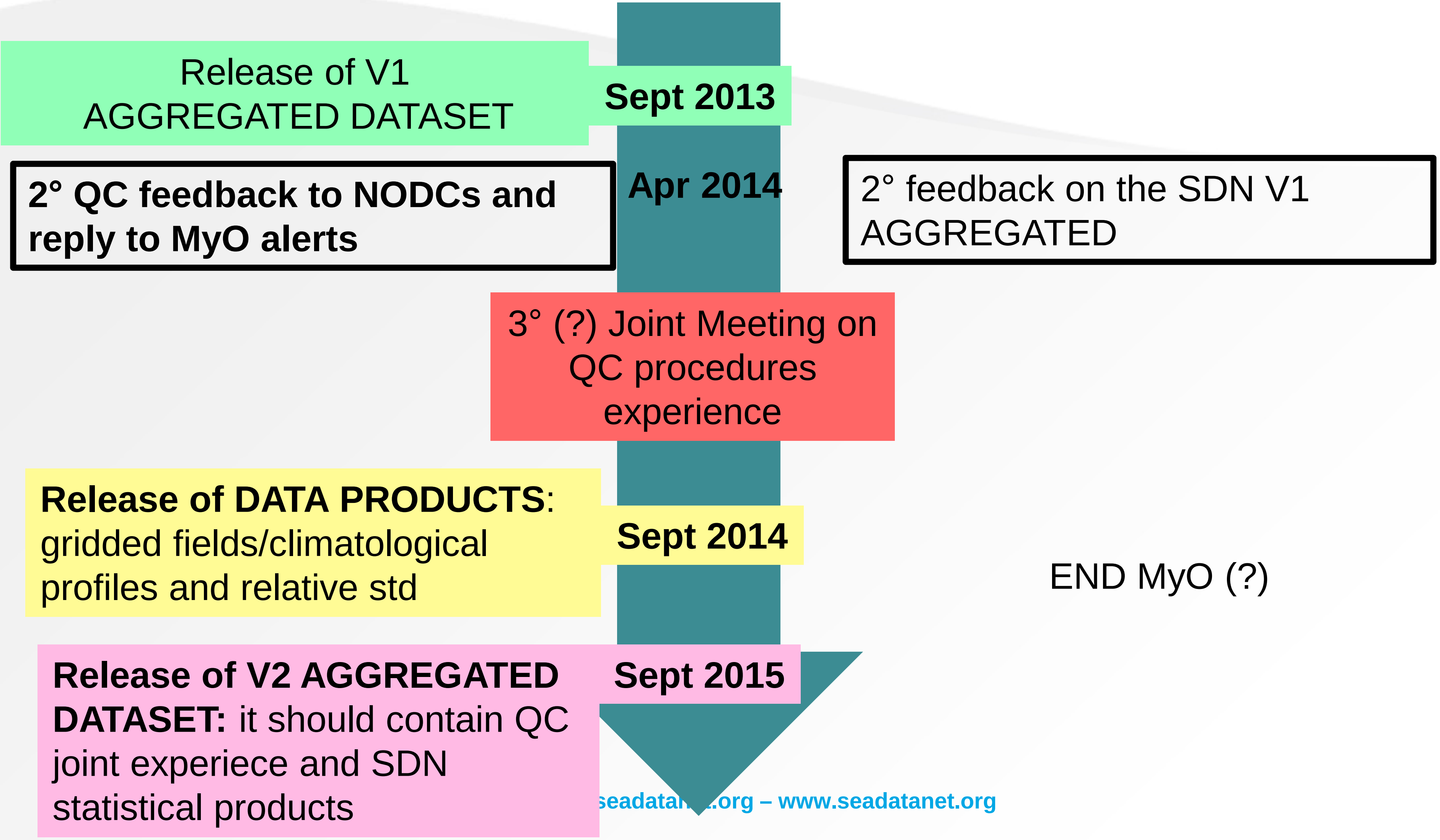
To be done ASAP

1. Finalize and harmonize the reports to include a detailed descriptions of: (a) analyses performed; (b) actions to be taken; (c) advices on how to use data
2. Send reports and lists of anomalies (with priorities) to data providers with a request to make corrections of original data
3. TBD: update of V1 before August

Future Work Plan

SDN2

MyO2



Release of V1 AGGREGATED DATASET

Sept 2013

2° QC feedback to NODCs and reply to MyO alerts

Apr 2014

2° feedback on the SDN V1 AGGREGATED

3° (?) Joint Meeting on QC procedures experience

Release of DATA PRODUCTS: gridded fields/climatological profiles and relative std

Sept 2014

END MyO (?)

Release of V2 AGGREGATED DATASET: it should contain QC joint experiece and SDN statistical products

Sept 2015