SeaDataNet2-MyOcean2
2nd Joint Meeting on Products

Cork, 15th April 2013

The first objective of this half day meeting was to get a first feedback from MyOcean In Situ TAC partners on the Quality of the data provided by SeaDataNet and on the issues regarding the present interfaces. The second objective was to inform each project of their respective plans for 2013.

The agenda of the meeting was :

- 15h00-1520 : Summary of last Meeting agreement and revised MyOcean-INSTAC plans for V3 and V4 MyOcean (S Pouliquen)
- 15h20- 15h50 : From SDN side feedback on dataset aggregation (what worked well what have been the difficulties) Simona with SDN WP10 contributors
- 15h50-16h20: From MyOcean feedback on the weaknesses of the SDN dataset, summary on how we use the data for V3, Requirement for next version
- 16h45- 18h00 : discussion and elaboration of a plan for V4

I. Progress on MyOcean side

S Pouliquen reminded the participants on agreement decided at the Rhodes meeting last September. She presented the progress made on the MyOcean side in the integration of historical data both from EuroGOOS ROOS partners and from SeaDatanet. Due to the delay in providing the SeaDataNet input, the validation of the historical product was delayed on MyOcean side and focus was made on aggregating the data from differences sources , ingesting the data from SeaDataNet rapidly after it's availability and getting rid of the duplicates between the different data streams.

The regional products are built in two steps: first, for MyOcean V3, the focus is in the aggregation of the data from the ROOS providers and the SeaDataNet National Data Centres, removing duplicates and converting all data in the same format with the same QC flags. Whenever it's possible, all the parameters measured by a platform are aggregated even if the scientific validation will only be performed of the Temperature and Salinity parameters.

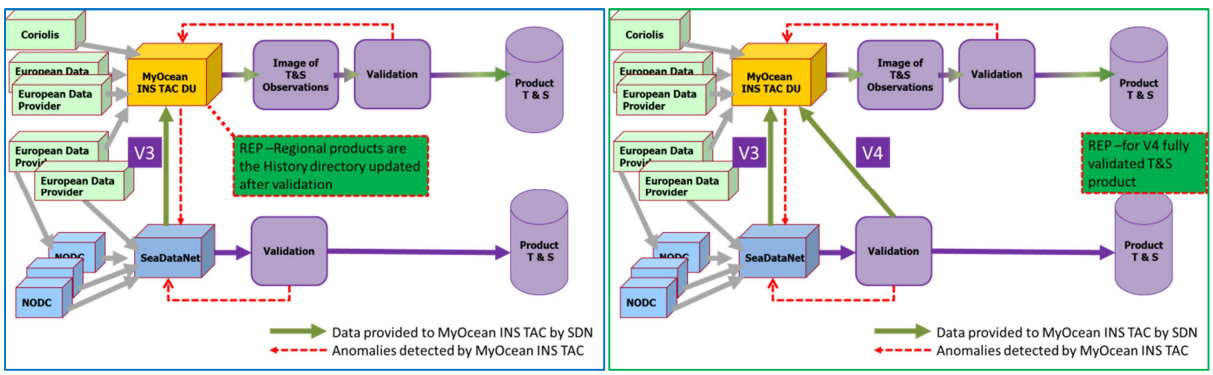

Figure 1 Building V3 and V4 Reprocessed T&S in situ products

II. Progress on SeaDataNet side

S. Simoncelli presented the process on SeaDataNet side to elaborate the first version of the aggregated dataset that was provided to MyOcean in February-March 2013.

She first revised the common time schedule (shown in Figure 2) defined in Rhode at the First Joint Meeting in September 2012, reminding that on day 10th of April 2013 the Regional Coordinators met to discuss about the first phase of the Quality Control (QC) process implemented on the first "raw aggregated datasets" released to the INS-TAC. Then she focused on the specific aims of this Second Joint Meeting, which were:

1. to feedback to the INS-TAC about the aggregation procedure and describe the preliminary quality control analysis applied;
2. to discuss with the INS-TAC about their QC results and about the possible strategies to follow in order to improve the overall quality of the aggregated dataset;
3. to define what to report to the SeaDataNet data providers (NODCs) asking to take actions on the data within the infrastructure. This methodology has been proposed to correct the data/metadata at the source and to facilitate the future upgrades of the joint products.

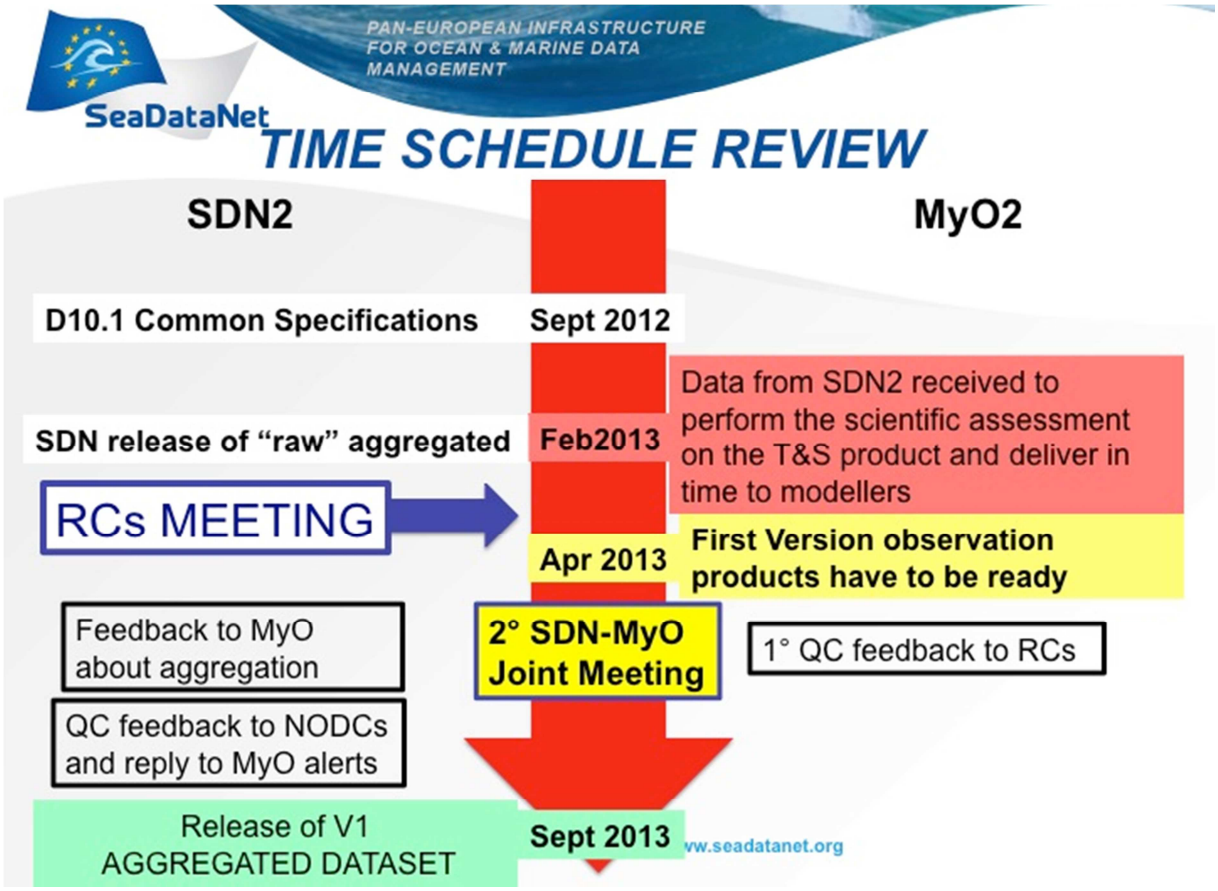

Figure 2 SeaDataNet WP10 and MyOcean INS_TAC common time schedule defined at the First Joint Meeting in Rhode on the 18th of September 2013.

S.Simoncelli then presented the **harvesting procedure of Temperature and Salinity files by CDI Robot** developed at **MARIS** (*D. Shaap*). The CDI Robot used the CDI Data Discovery and Access Service to query, shop and retrieve data sets from the distributed data centres in an automatic way. The query searched for all data sets with T&S and for which the access restriction was Unrestricted or under 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 procedure was used to test and tune the performance of the RSM – DM process and find the optimum data requests. All data requests were administered in the RSM and the tuning resulted in a slicing factor of 500 data sets per cycle of 10 minutes, which could be handled by all connected data centres. RSM keeps track of all data requests and repeats data requests in case of disturbances at DM level. Robot harvesting and tuning of the shopping system run from mid December 2012 to mid January 2013 then all retrieved ODV files with the full CDI metadata as CSV file were stored on a DVD and sent to AWI (*R. Schlitzer*). The ODV files (more than 2 Millions of SDN data files in ODV format) contained in most cases not only T&S but also additional observations.

AWI (*R. Schlitzer*) did the aggregation of all data files into a single TS Data Collection using SDN Importer of ODV 4.5.3. The aggregation of the many original temperature and salinity variables into single T and S variables (using “Aggregated Derived Variables” ODV option) has been done in 9 pieces of about 250,000 files each, and then re-combined. The logs of problem files were analyzed and sent to the coordinator and the data centers for fixing. It followed the creation of regional and 1900-2012 subsets and distribution to SDN regional groups.

S. Simoncelli then described duplicate analysis conducted within SeaDataNet Infrastructure between October and December 2012, before the harvesting process of Temperature and Salinity files. The Duplicates Implementation Plan was prepared by HCMR (*S.Iona)*. This plan was based on the duplicates checks conducted with ODV for the 6 SDN regions in September 2012. The implementation plan was prepared and sent to all SDN partners on early October 2012, 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 were few missing cases for partners who didn’t finish to process their data sets. A total of 60866 potential duplicates were identified (see Tab.1) among which (see Figure 3) 64% had to be updated (unknown, wrong, missing information), 6% had to be deleted, 29% had to be kept invariant (time-space differences less the threshold values, different measurement methods), 1% had to be replaced.

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

Tab. 1 Results of the duplicate analysis conducted within SeaDataNet Infrastructure between October and December 2012, thus before the harvesting process of Temperature and Salinity files.

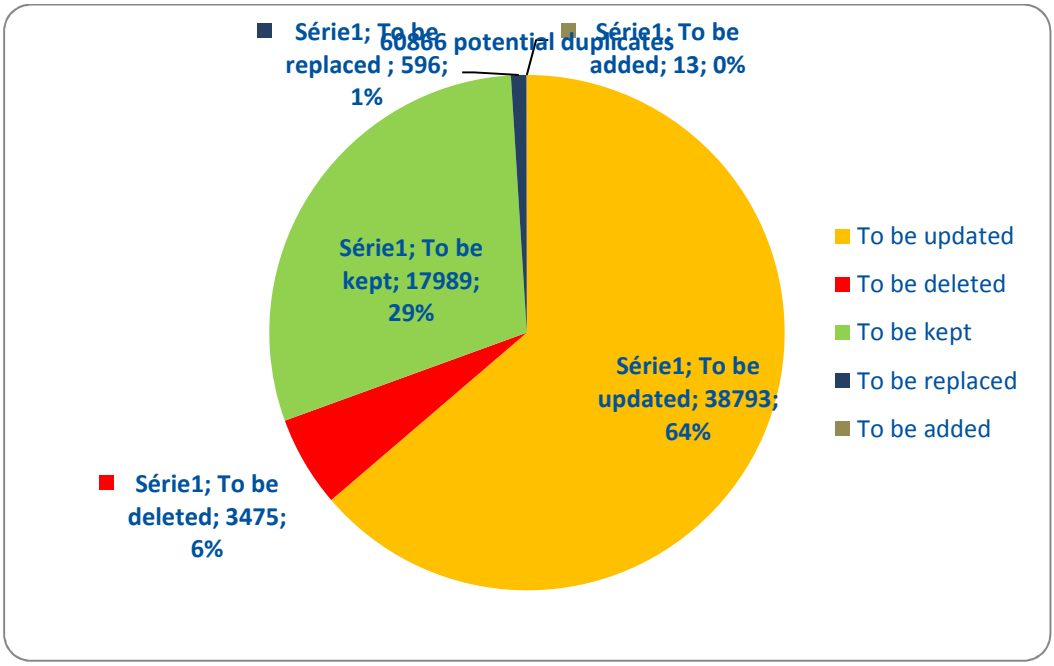

Figure 3 Summary of the actions that have been taken from the data providers (NODCs) on the 60866 potential duplicates identified within SeaDataNet infrastructure.

Guidelines will be sent to all SDN partners to avoid similar cases in the future, in the meantime a white list of the cleaned and checked CDIs has been prepared in order to check new entries in the CDI central catalogue to avoid future duplicates.

During the preparation of the aggregated dataset, more than 14 000 files were rejected because ODV was not standard or not SDN standard and ODV files had to be corrected. The list of errors was sent (*M.Fichaut*) to 33 data centres, among them 5 are not SDN partners (mid Feb2013). Not all these errors could have been fixed before data delivery to MyOcean.

Regional Coordinators, after a preliminary and basic QC through ODV software, released the "raw" aggregated temperature and salinity collections, covering the time period 1990-2012, to the regional responsible of the INS-TAC. The WP leader provided to the colleagues some guidelines and a report template in order to harmonize the work in progress. The outcome was presented first by the leader to the Steering Committee in Paris on March 27th 2013 in order to share the work done and receive feedback and advices on the developments. **A shared decision between WP10 leader and the steering committee was that RCs would not modify the data or the QC flags but it would define procedures to report to data providers in order to facilitate the update procedure and to**

progressively improve the overall quality of the infrastructure, consistently with what stated in SDN DoW.

The outcome of the preliminary QC analysis is:

1. T-S datasets require QC analysis regardless their QC flag to identify anomalies and possible solutions
2. Statistics about QC flags are necessary to monitor the overall quality of the database and its improvement
3. QC reports need to be further harmonized including TS scatter plots of the entire data set, of data having QC flags equal to 1 (good) and 2 (probably good), those with QC flags equal to 0 (not checked).
4. Visual control of scatter-plots is necessary to identify wrong profiles, outliers and spikes
5. Range check analysis is needed to save automatically a list of outliers.
6. Stations falling on land have to be identified and reported to the NODCs.
7. Missing data or Metadata have to be identified and reported to the NODCs.
8. Data providers (NODCs) with most problem data will be identified and solicited to act.

On the 10th of April 2013 the Regional Coordinators met in a WebEX online Conference with the aim to share their work, comment the results and find common strategies to feedback both the NODCs and MyOcean on the validation of the aggregated datasets. **The RCs decided to have a common strategy for future QC analysis and to refine QC procedures adding sub-regional QC (per areas & per depth) and stability check on density. They also aged on having a responsible person to coordinate the comunication between NODCs – RCs - MyO INSTAC. This person is *Christine Coatanoan* (Ifremer) with the help of *M. Fichaut* and *S. Iona (HCMR).*** The RCs agreed on identifying a list of priority actions, on the base also of MyOcean feedbacks, to be taken from the NODCs.

In conclusion *S.Simoncelli* presented the desired future efforts. The RCs 1 will:

1. finalize and harmonize the reports on the QC assessment of the regional T&S data collections to include a detailed descriptions of: (a) the analyses performed on the data collections; (b) the actions to be taken at the NODCs level; (c) advices on how to use data from a user prospective;
2. evaluate MyOcean INS-TAC feedback and provide reports and lists of anomalies (with priorities) to data providers with a request to make corrections of original data;
3. suggest and evaluate in accordance with the Steering Committee an additional update of the T&S collection before the official V1 release due for September 2013.

She finally presented the time schedule of future activities, shown in Figure 4.

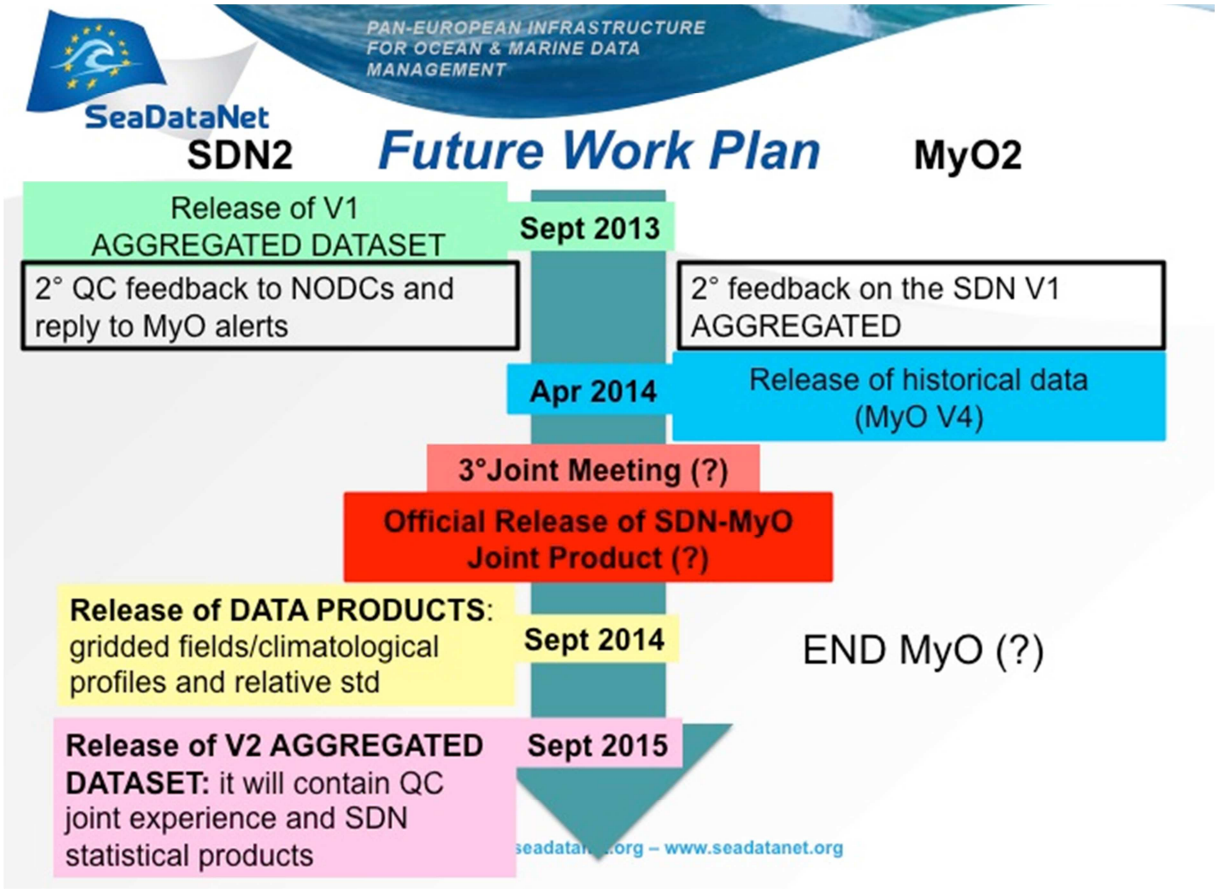

Figure 4 Schematic of the future activities of SeaDataNet WP10 and MyOcean INS_TAC

III. Feedback from MyOcean INSTAC on issues

Based on the IBI experience Thierry Carval reported the issues that are causing problems for an easy aggregation with the rest of the data managed by the INSTAC.

The main issue is that some important metadata have been lost in the aggregation process (i.e. **Platform CODE**, **Instrument type for XBT**,..) . Therefore, at the INSTAC meeting last February it was decided to insert in the V3 product only the data that were new and put aside the data that were detected as duplicates as it was difficult to know if these data were a better version of a data already managed by the INSTAC or another dataset. When INSTAC partners had already get a dataset directly from a ROOS partner, the SeaDataNET version was discarded as it had less metadata.

The **lack of Platform CODE** is critical because all MyOcean data are organised by platform.

Issues to be studied and if possible solved by SeaDataNet : In Red the crucial ones that should be studied in priority by SeaDataNet technical team

1. Many profiles have salinity but no temperature. Many profiles have no temperature and no salinity.
2. In the ODV file the vertical reference is immersion. In MyOcean we need the original measurement: pressure for CTDs.
3. For XBTs we need the WMO instrument type and the fall rate equation.

4. In addition to the cruise ID, we need a platform code for each profile. At present INSTAC create pseudo platform codes from the cruise ID but this is not satisfactory and will cause issues for the updates.
5. When a CTD is provided we would like to know whether CTD is calibrated or not, as it may be a criteria for the INSTAC to replace the same CDT received in RT *(S.Simoncelli: I recommend to keep the original profile and to add information, this would give the user the possibility to decide which data to consider)*. Modelers would like to know the correction applied on the CTD but this is not.
6. A fair number of data with QC 0 (no control) are mixed with QC 1 (good) either within a profile or in a series of profiles. We plan to change 0 into 1 after visual QC but they should be put at 1 by NODC.
7. some immersions are duplicated. If we perform the RTQC test on such profile we can detect them with the pressure increasing test but in theory we should not have to redo RTQC test of SeaDataNet data
8. Some longitude are in 0-360 instead of -180 and 180

Globally the data are good but a visual QC is useful (0,5% of suspicious data on the SeaDataNet dataset selected for IBI). In the IBI area, we plan to perform additional statistical test to check the consistency of the data over the area as well as climalological test to detect outliers.

INSTAC partners will generate feedback files according to the format defined in the MyOcean specification document. **The target is to provide the input to Christine Coatanoan by mid-May2013, to allow a feedback to NODC by end of May 2013.**

A feedback from SeaDataNet on the critical issues will be appreciated in the coming weeks so that the update (*S.Simoncelli: the update is not decided yet, the issue will be presented at the StComm and the decision will be taken soon*) planned in September is easier to handle.

Attendance

NO	FAMILY NAME	FIRST NAME	ORGANIZATION	COUNTRY
1.	Scory	Serge	RBINS-MUMM	Belgium
2.	Marinova	Veselka	Institute of Oceanology (IOBAS)	Bulgaria
3.	Carval	Thierry	IFREMER	France
4.	Coatanoan	Christine	IFREMER	France
5.	Pouliquen	Sylvie	IFREMER	France
6.	Gies	Tobias	BSH Hamburg	Germany
7.	Chalkiopoulos	Antonis	HCMR	Greece
8.	Perivoliotis	Leonidas	HCMR	Greece
9.	Simoncelli	Simona	INGV	Italy
10.	Tonani	Marina	INGV	Italy
11.	Lid Ringheim	Sjur	IMR	Norway
12.	Sagen	Helge	IMR	Norway
13.			IMR	Norway
14.	Jaccard	Pierre	NIVA	Norway
15.	Sorensen	Kai	NIVA	Norway
16.			NIVA	Norway
17.	Gorringe	Patrick	EuroGOOS	Sweden
18.	Hammarklint	Thomas	SMHI	Sweden
19.	Back	Orjan	SMHI	Sweden