Data Quality: why it is so important?
The experience of the regional products managers and recommendations

S. Scory, S. Simoncelli, C. Coatanoan, Ö. Bäck, S. Iona, V. Myroshnychenko
• SeaDataCloud data products and objectives
• SeaDataCloud Quality Check Strategy
• SeaDataCloud products’ timeline
• Quality Control procedure
• Data anomalies and data providers’ response
• The importance of full metadata record
• Unlock your data and set them free
• PIDocs and acknowledgment of data providers
• SeaDataCloud innovation
SeaDataCloud Data Products

SeaDataCloud aims at providing **data products** deriving from SeaDataNet infrastructure at **regional and global scale** to serve a diverse user community:

1. **Aggregated data sets for all the European marginal seas** → all historical temperature and salinity (1900 onwards) data harvested from the central CDI and validated by regional leaders

2. **Climatologies** → gridded fields obtained through a mapping technique (DIVA) and representing the climate of the ocean at both regional and global scale

3. **New data products** → multi-platform and multi-disciplinary approach combining both in situ (e.g. gliders, Argo, ships, drifters, fixed platforms) and remote sensed observations, Ocean Monitoring Indicators for tracking ocean mechanisms and/or climate modes and trends
Objectives

- **Improve the quality of the overall infrastructure content** through systematic quality assessment (every 2 years)
- **Develop new methods** to ensure quality, homogeneity and robust uncertainty measures in long-term time-series of data
- **Integrate external datasets** (Copernicus Marine Environment Monitoring Service, World Ocean Database) to increase temporal and spatial resolution and further improve products’ quality
- **Generate the best data products** to serve different user groups (operational oceanography, climate, marine environment, institutional, academia) adopting the most advanced methodologies
- **Increase user uptake** providing timely and reliable information of the full product generation process and its quality
Example application

Ocean Prediction

Atm Forcing

Ocean general circulation models (OGCMs)

Initial Condition

Data Assimilation

Forecast/analyses reanalyses
Ocean Prediction

Atm Forcing

Ocean General Circulation Models (OGCMs)

Data Assimilation

Initial Condition

Forecast/analyses reanalyses

Validation procedures rely on climatologies, gridded reconstructions, reprocessed time series of data

Reanalyses ➔ harmonized historical data collections
SDN2 project implemented and continuously refined a **Quality Control Strategy (QCS)** aiming at improving the quality of the database content and creating the best data products.

**Iterative approach** to facilitate the **upgrade** of the database and **versioning** of data products through:

- the release of new data collections at the end of each QCS loop
- the generation of derived climatological products after a certain time lag dedicated to data processing
Quality Check Strategy

Data providers have to timely analyze the list of anomalies and make the necessary corrections on the quality flags or the data format and update the CDI.

Regional products leaders compile a list of data anomalies and organize it per EDMO code. The list of anomalies is sent to the data providers.

A rapid feedback from the data providers guarantees:
- the timely generation of data products → increasing user confidence and awareness
- the upgrade of the database content → no mismatch among products and CDI service
* It is crucial to correct data anomalies (M26→Dec2018) and ingest all available data before next data harvesting (M30→Apr2019)
Quality Control procedure

The quality control work follows the best practices that were defined during the project SeaDataNet 2:

- **Checks of the data coverage, by sub-region when necessary** (distribution for T, S, TS couples), by time periods, by layers (distinction between surface, intermediate and bottom layers);
- TS scatter plots of the entire dataset: T versus Z, S versus Z, $\theta$S diagram with isopycnal levels for all the QF<3 (check the outliers and change the QF to 4); sometimes the outliers were the missing data values with not appropriate QF;
- By sub-region, scatter plot of observations with QF=1 (good) with a secondary plot showing the density;
- By sub-region, scatter plot of observations with QF=2 (probably good) with a secondary plot showing the density;
- Scatter plot observations with QF=0 (no quality check): only change the bad data with QF4;
- Identification of stations falling on land;
- Identification of stations having unreal depth (depth values<0);
- The most useful and powerful quality control used was visual inspection of subsets of data in ODV to discover spikes, outliers, unstable profiles and stations on land.
Quality Control procedure

- QF statistics on correction

<table>
<thead>
<tr>
<th></th>
<th>BEFORE CORRECTION (%)</th>
<th></th>
<th>AFTER CORRECTION (%)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>QC0</td>
<td>QC1</td>
<td>QC2</td>
<td>QC3-4 (5-9)</td>
</tr>
<tr>
<td>ARTIC SEA</td>
<td>0.66</td>
<td>98.99</td>
<td>0.01</td>
<td>0.34</td>
</tr>
<tr>
<td>TEMPERATURE</td>
<td>0.68</td>
<td>98.35</td>
<td>0.01</td>
<td>0.96</td>
</tr>
<tr>
<td>SALINITY</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BALTIC SEA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TEMPERATURE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SALINITY</td>
<td>11.96</td>
<td>83.87</td>
<td>2.88</td>
<td>1.28</td>
</tr>
<tr>
<td>BLACK SEA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TEMPERATURE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SALINITY</td>
<td>11.6</td>
<td>91.9</td>
<td>4.62</td>
<td>1.89</td>
</tr>
<tr>
<td>NORTH ATLANTIC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TEMPERATURE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SALINITY</td>
<td>2.38</td>
<td>95.39</td>
<td>0.23</td>
<td>2</td>
</tr>
<tr>
<td>NORTH SEA (discrete)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TEMPERATURE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SALINITY</td>
<td>2.7</td>
<td>96.9</td>
<td>0.3</td>
<td>0.04</td>
</tr>
<tr>
<td>MED SEA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TEMPERATURE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SALINITY</td>
<td>4.5</td>
<td>94.6</td>
<td>0.9</td>
<td>0.04</td>
</tr>
</tbody>
</table>
Issues with metadata

• Minimal set of metadata → almost useless
• Wrong “measuring area type” (one single measurement defined as “curve”)
• Moorings defined as “profiles” instead of time series” (particular case of Thermistor chains in one CDI → “Time series of times series”?)
• Vertical resolution given in “minutes”
• Station on land
On land
Data anomalies
Out of range with QF1
Bad values – Zero reported instead of default values
Downcasts and upcasts together
Sensor not stabilized
Unstable profiles

In situ Density Anomaly [kg/m$^3$]

ITS-90 water temperature [degrees C]

Water body salinity [per mille]
Sensor issues
Spike (big and small)
Artefact of ODV aggregation

One CDI_ID with 17 stations
## Artefact of ODV aggregation

### Aggregation procedure with P35

**Step 1** ➔ SDN QC flags

<table>
<thead>
<tr>
<th>PRES</th>
<th>PSAL</th>
<th>SSAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1774.0</td>
<td>1</td>
<td>35.086</td>
</tr>
<tr>
<td>1775.8</td>
<td>1</td>
<td>35.083</td>
</tr>
<tr>
<td>1777.0</td>
<td>1</td>
<td>35.081</td>
</tr>
<tr>
<td>1778.0</td>
<td>1</td>
<td>32.123</td>
</tr>
</tbody>
</table>

**Step 2** ➔ ODV QC flags

<table>
<thead>
<tr>
<th>PRES</th>
<th>PSAL</th>
<th>SSAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1774.0</td>
<td>0</td>
<td>35.086</td>
</tr>
<tr>
<td>1775.8</td>
<td>0</td>
<td>35.083</td>
</tr>
<tr>
<td>1777.0</td>
<td>0</td>
<td>35.081</td>
</tr>
<tr>
<td>1778.0</td>
<td>0</td>
<td>32.123</td>
</tr>
</tbody>
</table>

**Step 3** ➔ Median value for aggregation

<table>
<thead>
<tr>
<th>PRES</th>
<th>Salinity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1774.0</td>
<td>35.0865</td>
</tr>
<tr>
<td>1775.8</td>
<td>35.0835</td>
</tr>
<tr>
<td>1777.0</td>
<td>35.081</td>
</tr>
<tr>
<td>1778.0</td>
<td>32.123</td>
</tr>
</tbody>
</table>

**Step 4** ➔ Keep the worst ODV flag

<table>
<thead>
<tr>
<th>PRES</th>
<th>Salinity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1774.0</td>
<td>0</td>
</tr>
<tr>
<td>1775.8</td>
<td>0</td>
</tr>
<tr>
<td>1777.0</td>
<td>0</td>
</tr>
<tr>
<td>1778.0</td>
<td>0</td>
</tr>
</tbody>
</table>

**Step 5** ➔ Back to SDN flag

<table>
<thead>
<tr>
<th>PRES</th>
<th>Salinity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1774.0</td>
<td>1</td>
</tr>
<tr>
<td>1775.8</td>
<td>1</td>
</tr>
<tr>
<td>1777.0</td>
<td>1</td>
</tr>
<tr>
<td>1778.0</td>
<td>1</td>
</tr>
</tbody>
</table>
From EMODnet-Chemistry

• Two parameters with the same user label (name) and different units
  – labels were renamed
Values 999 with flag 1

Flag changed to 4

Negative values

Flag changed to 4
Wrong values with flag 1

Mediter. max depth = 5121 m
– depth flag changed to 4
– Parameter flag changed to 4
Wrong units

No Qced data

Flag changed to 4

Flags changed to 1
Data anomalies

Examples of the various anomalies:

• Format issues: Several missing data values 999.999 or 99.999 or 9.999 or -999.00 or 99.00 and QF0 => missing data can have several values (rules of NODC) but the QF should be 9 (then in ODV: values will be empty and QF9)
• QF 0 → must disappear!
• Raw CTDs
• Down- & upcasts together, non-stabilized sensors
• Missing data badly flagged
• Out of range
• artefacts generated by ODV aggregation
Feedback to data providers

What is sent to each CDI partner?

- The anomalies list with LOCAL_CDI_ID EDMO_CODE PARAMETER_LEVEL and OLDQC NEWQC, sometimes with more explanation if necessary (doc)

LOCAL_CDI_ID EDMO_CODE PARAMETER_LEVEL_OLDQC_NEWQC

......
Feedback to data providers

What is expected from each CDI partner?

- The anomalies list updated with NODC comments following this table
- The list of updated CDI:

<table>
<thead>
<tr>
<th>LOCAL_CDI_ID</th>
<th>EDMO_CODE</th>
<th>PLATFORM_CODE=CRUISE</th>
</tr>
</thead>
<tbody>
<tr>
<td>FI3519910300_00050_H10</td>
<td>486</td>
<td>PRIMO-0 21/03</td>
</tr>
<tr>
<td>FI35199443000_25900_H10</td>
<td>486</td>
<td>MBP-FRONT 1994</td>
</tr>
<tr>
<td>FI35199502000_00870_H10</td>
<td>486</td>
<td>EUROMARGE</td>
</tr>
<tr>
<td>FI35199706000_00K010_H10</td>
<td>486</td>
<td>PELMED 97</td>
</tr>
<tr>
<td>FI35199845000_00260_H10</td>
<td>486</td>
<td>BIODYPAR 1</td>
</tr>
</tbody>
</table>

- A report with some informations:
  - List of errors and number
  - Details on why corrections have not been taken into consideration, etc......
Unlock your data and set them free!

• Importance of sharing data for knowledge advancement
• Make your restricted data → unrestricted
Goal: to associate to each product a **PIDoc** containing all the specifications about its:

- General characteristics (format, space-time coverage, resolution)
- Quality (validation methodology and results)
- Usability

**PIDoc** will have a DOI as well as the data products and both will be available through the SDC product catalogue

→ This would increase user confidence and uptake of SDC products

→ It would also provide details on how to reproduce the products in the VRE where data and tools will be available
Goal: to associate to each product a **PIDoc** containing all the specifications about its:

- General characteristics (format, space-time coverage, resolution)
- Quality (validation methodology and results)
- Usability

**PIDoc** will have a **DOI** as well as the data products and both will be available through the SDC product catalogue

→ This would increase user confidence and uptake of SDC products

→ It would also provide details on how to reproduce the products in the VRE where data and tools will be available

Each **PIDoc** contains the full list of data distributors and originators

→ acknowledgment of your effort!
The implementation of the cloud environment will optimize and automate the QCS at the central level assuring a continuous monitoring of the database content and quality, together with the possibility of generating database snapshots on a regular basis and allowing data products versioning.
Reactions?
The **ingestion of new data types** (HF radar, glider data) and the **integration of external data sets** are fundamental actions for the creation of appropriate observational data products as demanded by the user community.
Virtual Research Environment

New Data Products

Standard Data Products
historical data collections
climatologies

data subsets
Virtual Research Environment

• The positive impact of VRE on data providers (i.e. automatic management of data anomalies)

• See following slide (to get a message each time a QF is defined as doubtful or bad, correct and charge automatically the corrected data)
How to improve QCS? to an automated way

2nd release: possibility to get only new, updated profiles? List of removed profiles

Work on a unique subset

Message for NODC, in the VRE message window?

LIST OF STATIONS WITH CORRECTION ON QC

Central CDI catalogue

Data Harvesting

File and Parameter Aggregation

DATA

Message for NODC, in the VRE message window?
CONCLUSIONS

SeaDataCloud work plan on data products is very ambitious and our success is dependent from data availability and technical developments related to the cloud virtual research environment.

- More data → highest product quality and increased knowledge
- VRE will allow a fastest access to the data and the tools that will be shared