2nd Plenary meeting, Lucca, Italy, 26-27 Sept., 2013

SeaDataNet

PAN-EUROPEAN INFRASTRUCTURE FOR OCEAN & MARINE DATA MANAGEMENT

WP5 – Data: CDI directory

Results of the duplicate checks

Sissy Iona, S. Balopoulou (HCMR)
Reiner Schlitzer (AWI)

Objective

- To clean the CDI central archive from duplicates
  - To deliver to its users data of high quality
    - duplicates bias the analysis
  - To deliver to MyOcean Project a clean collection of observational data set for the production of the first version of climatology

Methodolgy

- Lists of **potential duplicates** were extracted from the CDI central catalogue for the 6 SeaDataNet regions
- duplicates detection **by ODV** using:
  - station longitude and latitude agree within 0.001° (ca. 100 m) and
  - station date/time agree within 3 minutes (to avoid time series data with 5 minute cycle as duplicates) and
  - the CDI instrument type agrees (to avoid CTD and bottle data as duplicates)

Methodology

- An **implementation plan** was prepared and sent to all SDN partners (and later to UBSS, GeoSeas) asking for:
  - identification of duplicates
  - cleaning of the local data sets (by deleting, updating, replacing, keeping)
  - provide detailed explanations for their actions

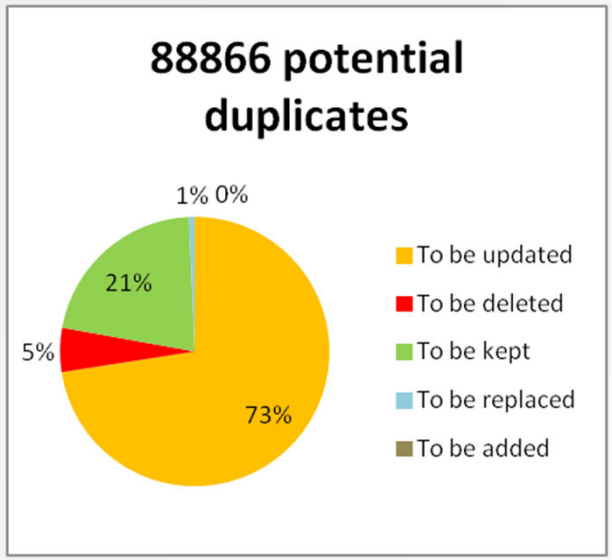
Cases of potential duplicates

Based on the archive content and partner's feedback, the potential duplicates were categorized in seven major cases:

- Identical copies of the same dataset
- Different distributors submitted subsets (in terms of station_id) of the same data set
- One distributor submitted different parts of the same data set
- One data set of the same instrument split by parameter
- Unknown, wrong, missing Information in data sets or false data sets
- Replicates
- Not duplicates

Results

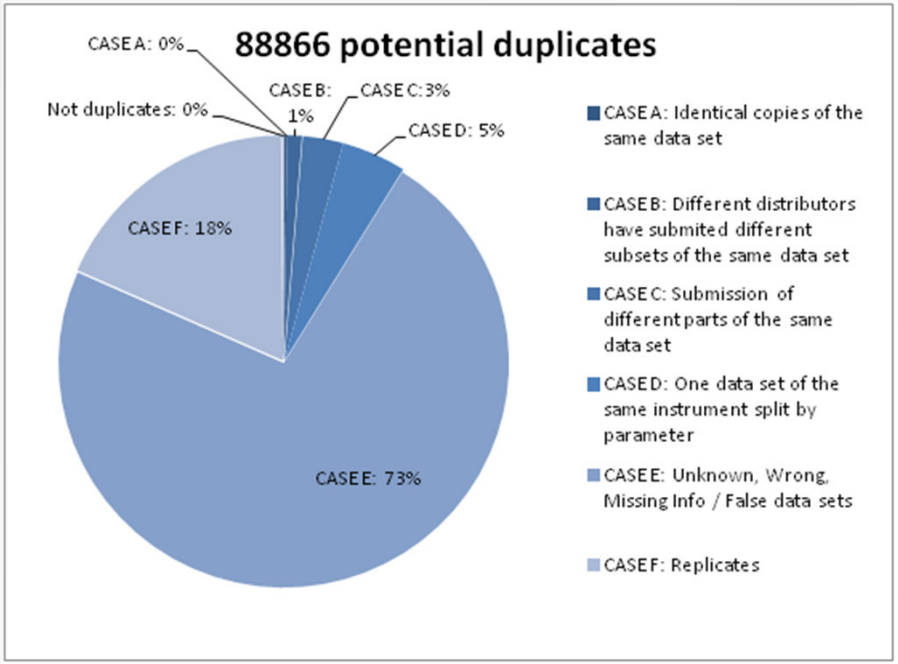
27/29 Partners	Potential duplicates	To be updated	To be deleted	To be kept	To be replaced	To be added
Total	88866	64477	4702	19078	596	13



- 5% were real duplicates and deleted
- the majority (73%) needed correction
- 21% were not duplicates and remained as they were

Results per case

CASE A: Identical copies of the same data set	CASE B: Different distributors have submitted different subsets of the same data set	CASE C: Submission of different parts of the same data set	CASE D: One data set of the same instrument split by parameter	CASE E: Unknown, Wrong, Missing Info / False data sets	CASE F: Replicates	Not Duplicates	Total
224	926	2587	4120	64645	16239	125	88866
0%	1%	3%	5%	73%	18%	0%	100%



Conclusions and future control

- The main reason for the existence of potential duplicates within the CDI catalogue was the missing or the wrong CDI time information, thus
  - Guidelines for preventing CDI duplicates generation before import to the central CDI catalogue (to be integrated with the manual for the management of the CDI data)

Central control for new submissions

- Andaas part of the upgrading of the whole CDI system to CDI ISO1939, a duplicate detection (same algorithm than the one used by ODV), has been integrated in the import validation process
  - New entries that might be potential duplicates will be flagged as suspicious, data providers will be notified and will be removed