SeaDataNet: What's up in the cloud?

Michèle Fichaut, IFREMER (France), Michele.fichaut@ifremer.fr
Dick Schaap, MARIS (The Netherlands), dick@maris.nl
on behalf of the SeaDataCloud consortium

Access to marine data is of vital importance for marine research and a key issue for various studies, from the climate change prediction to off shore engineering. But, the marine observing systems are highly fragmented: for example, in the countries bordering the European seas, more than 600 scientific data collecting laboratories from governmental organisations and private industry are collecting marine data by using various sensors on board of research vessels, submarines, fixed and drifting platforms, airplanes and satellites, to measure physical, geophysical, geological, biological and chemical parameters, biological species etc. The collected data used to be neither easily accessible, nor standardized. They were not always validated and their security and availability not assured.

Therefore the standardised SeaDataNet infrastructure was established for managing the large and diverse data sets collected by the oceanographic fleets and the automatic observation systems. It was implemented in several phases during the SeaDataNet project (2006-2011), grant agreement 026212, EU Sixth Framework Programme, the SeaDataNet 2 project (2011-2015), grant agreement 283607, EU Seventh Framework Programme and the current SeaDataCloud H2020 European project (under grant agreement 730960).

SeaDataNet contributes to build research excellence in Europe by networking and enhancing the currently existing infrastructures, which are the National Oceanographic Data Centres (NODC) or data focal points of 34 countries, active in data collection. The networking of these professional data centres, in a unique SeaDataNet virtual data management system provides integrated data sets of standardized quality on-line.

SeaDataCloud project, started in November 2016, aims at considerably advancing SeaDataNet Services and increasing their usage, adopting cloud and High Performance Computing technology for better performance, with the following objectives:
- improve discovery and access services for users and data providers,
- optimise connecting data providers and their data centres and data streams to the infrastructure,
- improve interoperability with other European and international networks to provide users overview and access to additional data sources and
- develop a Virtual Research Environment with tools for analysing data and generating and publishing data products.

The presentation will give an overview of the overall setting and background of the SeaDataNet infrastructure, the partnership and relations with other projects and will insist more particularly on the new challenges and the technical progress that have been achieved since the start of SeaDataCloud project, involving different partners of the Technical Task Group:

- The new workflow (see Figure 1) and new service components that are developed using the EUDAT cloud: replication of data from the data centres connected to SeaDataNet infrastructure, quality checks on data file formats, data file integrity and compliancy with metadata, duplicate detection… (MARIS, IFREMER, CINECA, DKRZ, GRNET, CSC, STFC)
- The expansion and extra services on the SeaDataNet common vocabularies (BODC)

- The implementation of the linked data concept in order to improve the connections between SeaDataNet catalogues and improve the user's search functions (MI),
- The development of a Sensor Web Enablement (SWE) service for facilitating the publication of real-time sensor data and their historical time-series through interoperable standards (52°North),
- The definition of standard data format for new data types such as flow cytometer data (CNRS-MIO, BODC, IFREMER),
- The transformation of SeaDataNet formats into INSPIRE data standards positively analysed (BODC, SYKE, OGS),
- The development of a Virtual Appliance (VA) to ease the installation of the SeaDataNet componant at data centres (ENEA)
- The specification of a Virtual Research Environment (VRE) offering users the possibility to work online with high capacity and performance for big data and to configure virtual workspaces for individuals or group with dedicated pools of data (MARIS, IFREMER, AWI, CINECA, DKRZ, GRNET, CSC, STFC, ULiège, VLIZ, SYKE, Deltares)

1: Data Centres provides CDIs (Common Data Index) to SeaDataNet

2: Import Manager (IM) runs metadata Quality Checks

3: IM triggers the Replication Manager (RM)

4 : RM upload a copy of the data in EUDAT cloud through B2STAGE. The uploaded data are stored in the import area of the B2SAFE

5: B2STAGE verifies authentication and authorisation of the data centre through B2ACCESS

6: IM orders and interacts with B2HOST through B2STAGE to submit the data sets to the series of quality checks

7: B2STAGE notifies the IM on the results of each quality checks as there might be reports which require decisions by the data centres

8: IM is the pilot of the import process and communicates via B2STAGE. Steps 7 and 8 are repeated for each consecutive application in the workflow

9: Metadata and related data with all checks performed successfully are moved to the production area of CDI catalogue and EUDAT cloud
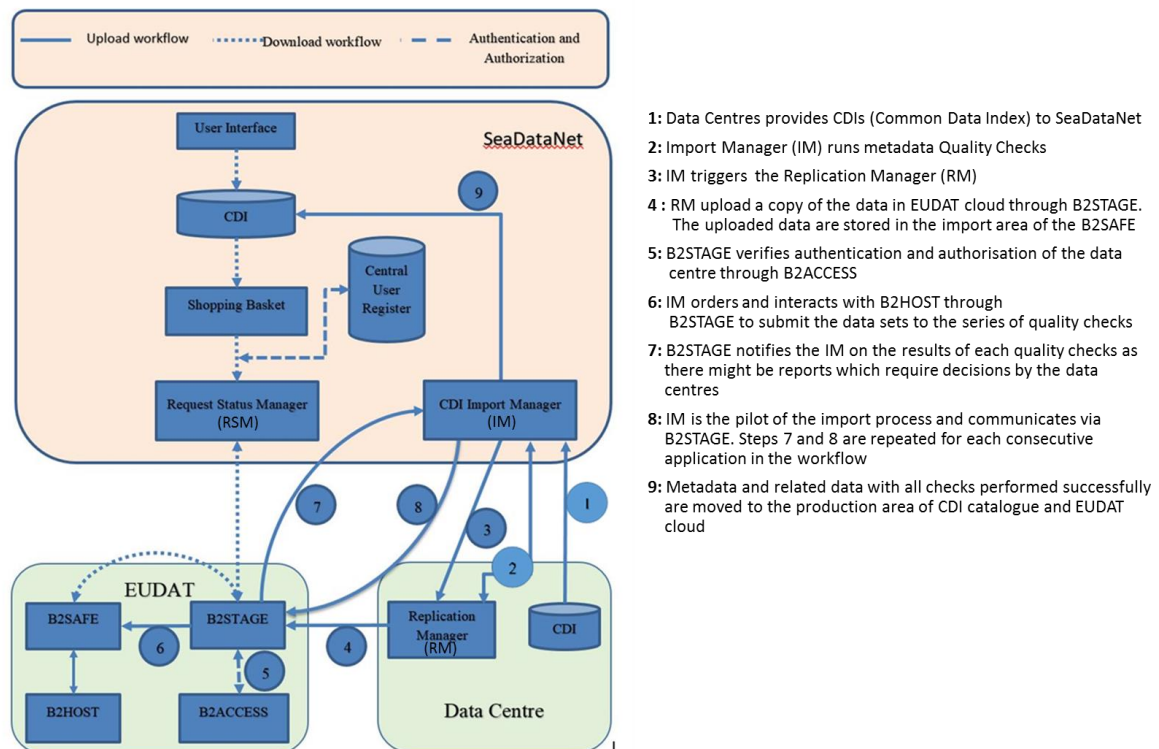
Figure 1: New workflow in SeaDataNet infrastructure, making use of EUDAT cloud and of the Replication Manager new SeaDataNet components

All these new developments are on-going, some of them are implemented in beta version others are already operational. The final objective is to provide benefits to SeaDataNet users and data managers, such as, for example:

- Improvement of the data downloading service (speed-up the performance, expand data discovery and ease of use of data access and downloading),
- Improvement of the data and metadata quality and coherence,
- The easier installation of the SeaDataNet connection "package' in the data centres.