Data management in Eurofleets+: the whole picture

Thomas Vandenberghe, Royal Belgian Institute for Natural Sciences (Belgium), tvandenberghe@naturalsciences.be
Susana Diez Tagarro, Consejo Superior de Investigaciones Científicas (Spain), sdiez@utm.csic.es
Dick Schaap, MARIS B.V. (The Netherlands), dick@maris.nl
Guillaume Clodic, IFREMER (France), Guillaume.Clodic@ifremer.fr
Juan Luis Ruiz, Consejo Superior de Investigaciones Científicas (Spain), jlruiz@utm.csic.es
Hong Minh Le, Royal Belgian Institute for Natural Sciences (Belgium), hmle@naturalsciences.be
Yvan Stojanov, Royal Belgian Institute for Natural Sciences (Belgium), ystojanov@naturalsciences.be
Christian Autermann, 52°North (Germany), c.autermann@52north.org
Simon Jirka , 52°North (Germany), s.jirka@52north.org

Eurofleets+ is a consortium of 42 research vessel operators aiming to provide access to ship-time for high-quality marine campaigns, including equipment and remote sampling access. From the start, the project has given data management a central place. This approach acknowledges the important drivers of efficient data management: a) broad acquisition by means of a data management plan, b) adequate transformation by software agents and c) integrating the exchange technology used by data repositories such as SeaDataCloud, all three designed to work together.

Eurofleets+ (EF+) is a 4-years H2020-funded project, and is currently in its second year. At this moment, no cruises have yet departed. For the cruise and dataset metadata funded by Eurofleets 2 (2013-2017), it has not always been apparent what their funding context was, let alone that a centralized view on the generated datasets was possible. For the Eurofleets+ proposal, the gaps in achieving this have been filled. For a better synergy with other aspects of the project, they have been separated into multiple work packages. Compared to Eurofleets 2, included in the description of work are a) the procurement of a data management plan (DMP) as a mandatory evaluation criterion, to assure data provision, and b) the assignment of dedicated data management organisations to assist principal investigators and vessel operators, to ensure the follow-up of the DMP and the data dissemination of EF+ cruises.

An additional reason to enforce DMPs is that it is a requirement of any H2020 project. Both the project and each individual cruise have a DMP. The cruise DMPs are managed on a forked DMP Roadmap web application (created by the UK Digital Curation Centre and the University of California Curation Center) and contains a number of questions adapted for EF+ from the H2020 Open Research Data Pilot.

The DMP website (http://dmp.ef-ears.eu) also provides the data management guidelines. These guidelines state the data workflow, from acquisition to dissemination. A distinction is made between en-route data and manual data. 'Manual' data (sample-derived) will be posted by the Principal Investigator on the EMODnet Data Ingestion Platform and data managed by three reference data centres, i.e. HCMR, OGS and BMDC. These will take care of the actual dissemination and promotion of both en-route and manual data by publishing the corresponding metadata in global directories (SeaDataNet and thence to EurOBIS, EMODnet, GEOSS, IOC-IODE portal) but also on a dedicated EF+ dataset catalogue, providing persistent links (DOIs) to the actual data, accessible through the project website and the "European Virtual Infrastructure in Ocean Research" portal (EVIOR). Specific attention is paid to 1) meteorological data, 2) "Essential Ocean Variables" (e.g. sea temperature, salinity, currents, oxygen, nutrients, carbon, plankton biomass,…), 3) 3.5 kHz or Chirp light seismic; and 4) multi-beam bathymetry, as these are underrepresented and have a high potential.

The main software agent is the Eurofleets Automated Reporting System (EARS) which provides software and services for en-route data acquisition, records cruise and event metadata, and transforms this metadata into the necessary European and global marine data standards. An optimized EARS "v2.5" will be distributed to vessel operators for use during the first few 2021 cruises (All 2020 cruises have been postponed due to the COVID-19 crisis). Version v3.0 is under development and will be released in April 2021.

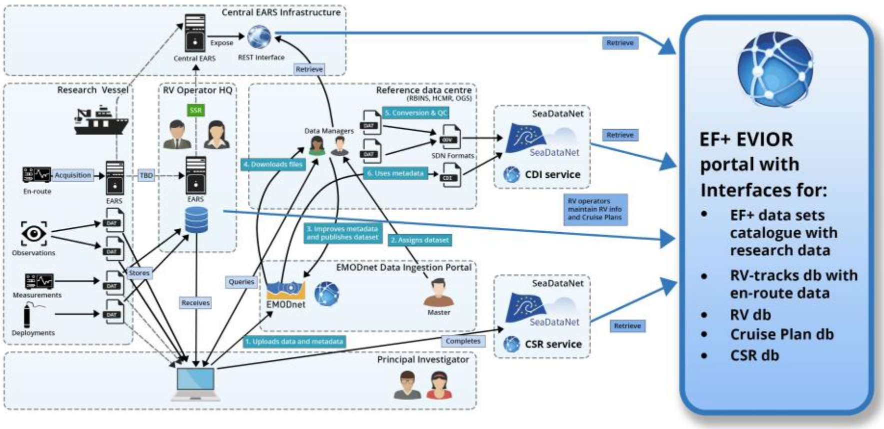

Figure 1: The complete project data workflow

The new EARS server distribution is based on docker and available on GitHub together with guidelines on installation. A docker image is a lightweight software package that combines the actual software plus the server environment. It is to be installed on the vessel, relies on TechSAS for data acquisition, and stores the metadata and data in the local EARS database, on the vessel. The metadata is transferred to shore by EARS, whenever the connection allows this, but the data transfer remains the responsibility of the vessel operator. The end goal is to let each RV operator have a 52°North Sensor Observation Service (SOS) installed on-shore as a central interoperability hub for acquisition data and event metadata. For operators without the possibility to install this SOS, a central SOS set up at CSIC will remain as a central datahub. The tracks and data of three Research Vessels are currently available on this datahub.

The EVIOR data portal connects to the SOS (GetObservation) to display the cruise tracks and primary en-route data (navigation, meteorology and thermosalinometry) as soon as it is made available by the vessel operator. The EVIOR portal and the operator's SOS will be the main interface for the data managers of the three reference data centres to retrieve the en-route data.

The EARS manual event database has been redesigned for EARS3, in order to contain all elements of a Cruise Summary Report, including references to P02, C77, the gml track and a summary of measurements. The java libraries to produce CSRs and Sensor ML are available on GitHub. In order to be able to express the device events in SensorML, a new BODC vocabulary (W11) has been created.

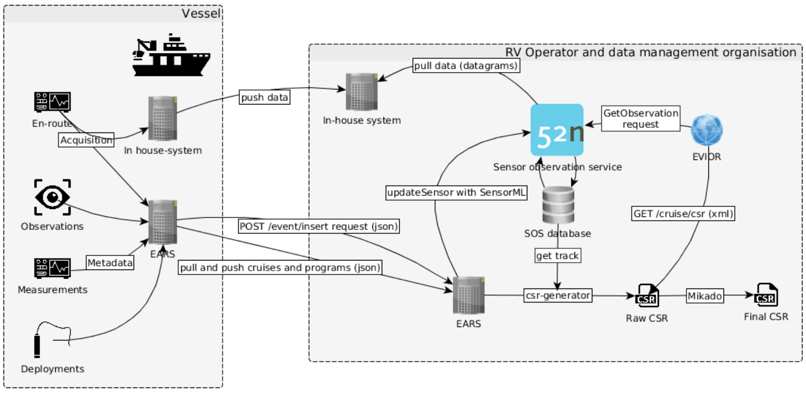

Figure 2: technical outline of the (meta)data flow