Real-Time Data Transfer and Management for the RV Belgica Using FROST OGC SensorThings API

De Ville de Goyet N.*, Van den Steen N.* & Vandenberghe T.*

Royal Belgian Institute of Natural Sciences

Our goal

Oceanographic **Research Vessels**, like the new RV Belgica, fulfil their scientific role by performing campaigns with specific scientific investigation and by continuously measuring the properties of the atmosphere and the water. Continuous data serve a **wide range of users**: the involved researchers, the global marine research community and private companies, the in-house dissemination tools and various Data infrastructures that RBINS constitutes and contributes to (e.g. GOSUD, INSPIRE, SeaDataNet,...). To meet the requirements of the different users a **fully operational sensor-to-client data flow** has been developed for all bound sensors that serves rich, standardized and quality-controlled data in near-real time.

Figure 1: RV Belgica. © Freire Shipyard

Design principles

The variety of user-profiles interested in our data is the best incentive to only store standardized data. We must be able to easily generate datasets for the different users with a minimum level of data transformations. For this reason, the **FAIR principles** have been followed during every step of the data architecture development. In previous projects, we developed an open-source solution for the provision of **INSPIRE-compliant metadata** and data using GeoNetwork and GeoServer (https://metadata.naturalsciences.be). These applications are closely following the Open Geospatial Consortium Standards (amongst others) and benefit from an active community. We decided to continue using the OGC Standards family with the Sensor Web Enablement (SWE) Standards (SensorML, SensorThings and Observations & Measurements).
We decided to select the dockerized FROST **SensorThings API** open-source implementation (https://sensors.naturalsciences.be/sta). The data model is simple, exposed using JSON and comes with powerful REST capabilities (time and spatial filtering, join operations, etc.).

Data Flow

Figure 1 shows the different components used for the automated data acquisition. Before the actual data dissemination to the different clients, the data needs to 1) get transferred from vessel to shore and imported into a replica of the vessel database, 2) be transferred to the SensorThings database with standardization and metadata enrichment, 3) be quality controlled, 4) be sub-sampled and transferred to the production database.

- Around **6 millions observations are recorded every day**. This includes meteorological, navigation, operational and physical/chemical parameters. A compressed selection of those parameters is sent via V-SAT connection every 10 minutes to an on-shore FTP server. Those data are imported into a replica of the Microsoft SQL vessel database, which is pruned regularly to save on storage. This database is vendor-specific and only serves the on-board data acquisition software but doesn't store the data in any standardized way. A second database is therefore needed for long-term storage.
- The georeference data are then imported into the full PostgreSQL database using the OGC SensorThings API endpoint. The database schema follows the SensorThings data model and is enriched with the necessary metadata (**NERC-controlled vocabularies** P01, P02, P06, L05, L22 and L20). The Fraunhofer FROST implementation provides a properties Jsonb column that can be easily used to incorporate NERC vocabularies.
- A Python application connects to the SensorThings API endpoint and performs **automated quality checks** on the location, range, gradient and spikes. The L20 quality flags are updated according to the test results. If the data passes the different tests, it gets a "probably good" flag. **The "good" flag is only granted when a human check is performed.**
- For performance reasons only a downsampled version of the data is exposed publicly. The time-series frequency is reduced to 10 minutes data, using only data that passed the quality checks, which is enough for most data clients. Full-frequency data is exposed only internally at RBINS.

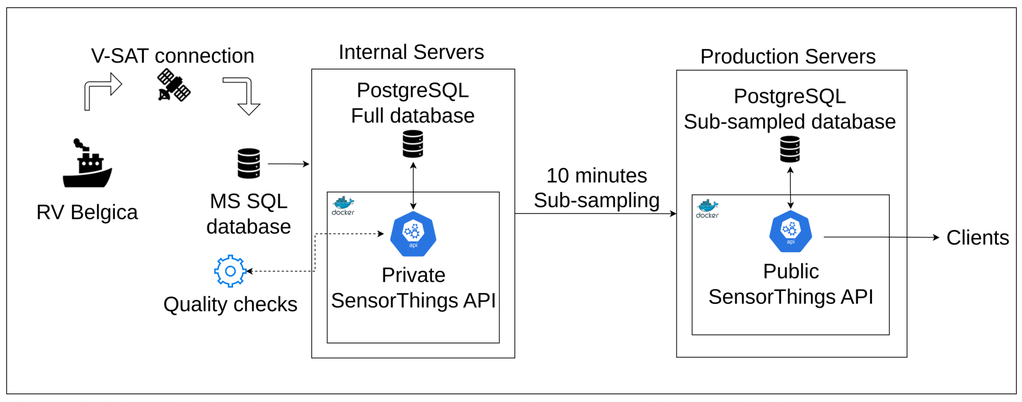

Figure 2: Schematic view of the different components of the automated data acquisition system.

Quality control

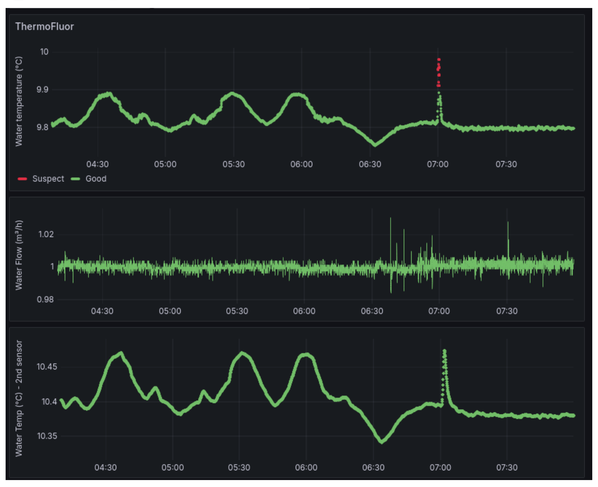

Figure 3: Data quality control. Illustration of the various steps involved in human verification following automated flagging.

The volume of data generated by modern, well-equipped research vessels is significant (e.g., 6 millions observations per day from 250 time-series for en-route data alone!). Data managers must be assisted with quality control through automated tools to save time for in-depth analysis of suspicious data. However, automating quality control is a challenging task. Observations can only be accurately labeled as good or bad when viewed from different perspectives.

- **Dependent/Independent Parameters**: For maintenance and safety reasons, most sensors are located inside the hull of the vessel. Water is brought to the sensors by a series of pumps, tubes, and valves. For validation, we must ensure that the water flow is high enough so that the water analyzed by the sensor is sufficiently renewed and its properties are not significantly altered (e.g., water temperature). All parameters are therefore dependent on the water flow. Furthermore, several parameters are computed based on other parameters (e.g., salinity, sound velocity). Primary parameters must be validated first before the computed parameters are checked.
- **Cross-Validation**: Some parameters are measured multiple times by different sensors at different locations. This should be used to analyze suspect trends. If a trend is present in all sensors, it can be considered possible. If only one sensor identifies it, it should be labeled as suspect.
- **Geo-Location dependencies**: Threshold checks should be performed with caution and with sufficiently wide boundaries, as water quality can significantly change when the vessel enters specific areas such as estuaries or high-latitude seas. Thresholds should always be analyzed in conjunction with gradients.

In the end, automated quality control is only part of the data management solution, allowing obvious errors to be excluded and highlighting suspicious data so the data manager's attention can be easily drawn to interesting events. **Figure 3** explains the process. The first graph shows suspicious water temperature data automatically flagged because of the steep gradient. The second graph shows that the water flow in the circuit was not disturbed and can't explain the spike. The third graph shows the water temperature measured by a second independent sensor. The spike data are therefore valid so data can be flagged as "good".

Conclusions

A fully automated near-real-time vessel-to-client data transfer has been implemented for the en-route data of the **new RV Belgica.** It includes the transfer itself, metadata enrichment, data standardization, quality checks and data dissemination. The infrastructure has been implemented using mainly existing **open-source solutions,** although some small components have been developed internally (data normalization and quality control). The quality control package is available on GitHub as open source (https://github.com/naturalsciences/qualityAssuranceTool). The FROST **OGC SensorThings API** proves to be a simple and reliable standard for the management and dissemination of sensor data, including various metadata (quality flags, geo-referencing, sensor information, etc.). One of the main issues is to ensure that the API performances for data retrieval (including filtering actions) are acceptable over time regardless of the database size. We decided to publicly expose only a sub-sampled version of the data for that purpose (https://odnature.naturalsciences.be/odanext/en).

natural sciences .be