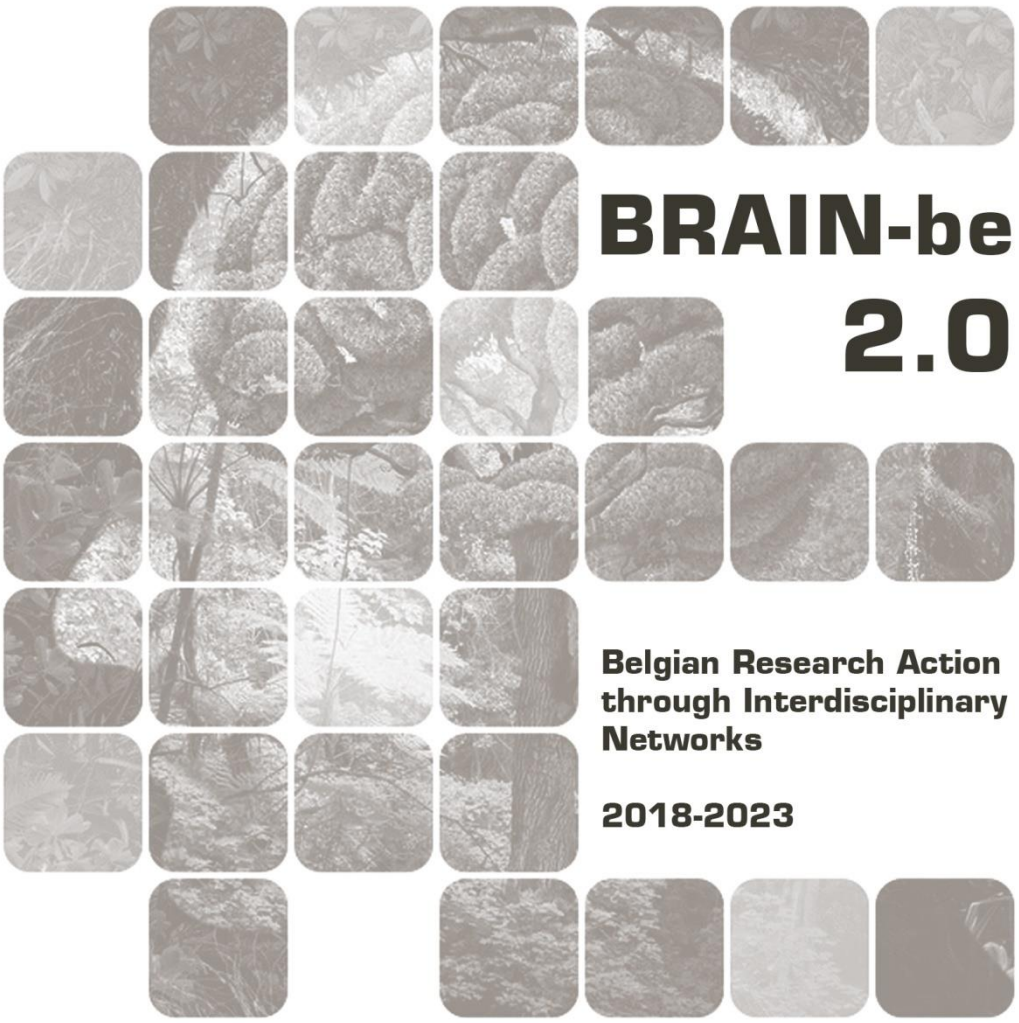


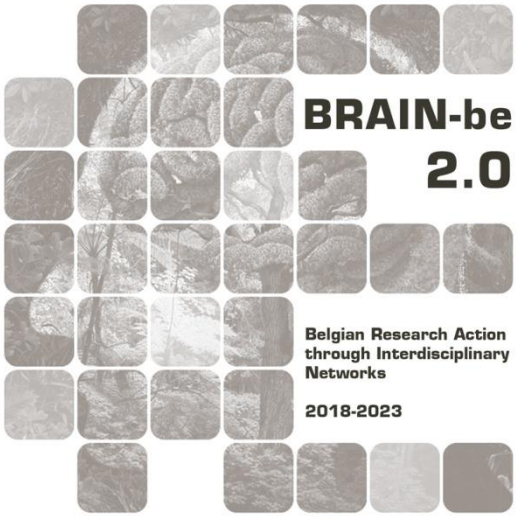
ODANext

Oceanographic data acquisition: the next age

de Ville de Goyet Nicolas (Institute of Natural Sciences)

Pillar 2: Heritage science

belspo .be



NETWORK PROJECT

ODANext

Oceanographic data acquisition: the next age

Contract - B2/202/P2/ODANext

FINAL REPORT

PROMOTORS: SCORY SERGE

AUTHORS: DE VILLE DE GOYET NICOLAS

VANDENBERGHE THOMAS

VAN DEN STEEN NILS

STOJANOV YVAN

YOUDJOU NABIL

VAN DEN BRANDEN REINHILDE

BACKERS JOAN

Published in 2024 by the Belgian Science Policy Office
WTCIII
Simon Bolivarlaan 30 bus 7
Boulevard Simon Bolivar 30 bte 7
B-1000 Brussels
Belgium
Tel: +32 (0)2 238 34 11
http://www.belspo.be
http://www.belspo.be/brain-be

Contact person: Georges Jamart
Tel: +32 (0)2 238 36 90

de Ville de Goyet, Nicolas. **Oceanographic data acquisition: the next age**. Final Report. Brussels: Belgian Science Policy Office 2024 – 23 p. (BRAIN-be 2.0 - (Belgian Research Action through Interdisciplinary Networks))

TABLE OF CONTENTS

ABSTRACT

The new Research Vessel Belgica represents a major advancement in Belgian maritime research, equipped with cutting-edge technologies and advanced scientific equipment. As a successor to the previous vessel, it introduces new challenges, particularly in managing vast amounts of data generated by its variety of sensors. The redesign of data systems is essential to fully utilize the vessel's capabilities, ensuring trustworthy data collection and analysis. This transformation is vital to meet modern IT standards and comply with European directives like INSPIRE and Open Data, which demand accessible, standardized, and open data. The ODANext project addressed these challenges by enhancing data acquisition workflows, preserving historical data, and establishing a durable infrastructure for data storage and dissemination. These improvements aim to elevate the RV Belgica's global research impact, facilitate efficient data sharing, and support scientific research, policy-making, and interdisciplinary collaboration.

1. INTRODUCTION

The newly commissioned Research Vessel Belgica (Figure 1) represents a significant leap forward in the Belgian maritime research capabilities. As the successor to the previous vessel of the same name, this state-of-the-art ship comes equipped with cutting-edge technologies, enhanced capabilities, and an extended operational range. The new RV Belgica is outfitted with advanced sensors and scientific equipment, enabling it to perform a wide array of complex research tasks in diverse marine environments. However, with these advancements comes the reality that many of the systems used on the former vessel are now outdated. Consequently, a comprehensive redesign of the data acquisition, management, and processing systems is imperative to harness the full potential of the RV Belgica. This redesign will ensure that the vessel's capabilities are fully utilized, allowing for precise data collection and analysis, and positioning the RV Belgica at the forefront of marine research.

Figure 1: The New RV Belgica

The introduction of the new RV Belgica also presents a multitude of challenges, particularly in managing the vast amounts of data generated by its operations. The scale of data collected by the vessel's numerous sensors and instruments are significantly greater than those of its predecessor. Effective data management is crucial not only to monitor and understand the behaviour of the vessel and its various components but also to transform the traditional methods of data handling. This transformation is essential to embrace the new linked-data era paradigm, which emphasizes interoperability, data sharing, and integration across different platforms and systems. Moreover, the fast-evolving IT environment, coupled with the need to comply with several European directives such as INSPIRE and Open Data, adds further complexity to this task. These directives mandate that data must be made accessible, standardized, and open, which requires the adoption of innovative approaches to data management. The volume of data generated by the new RV Belgica, which is an order of magnitude larger than that of the previous vessel, exacerbates these challenges, necessitating robust solutions for data storage, processing, and dissemination (Figure 2).

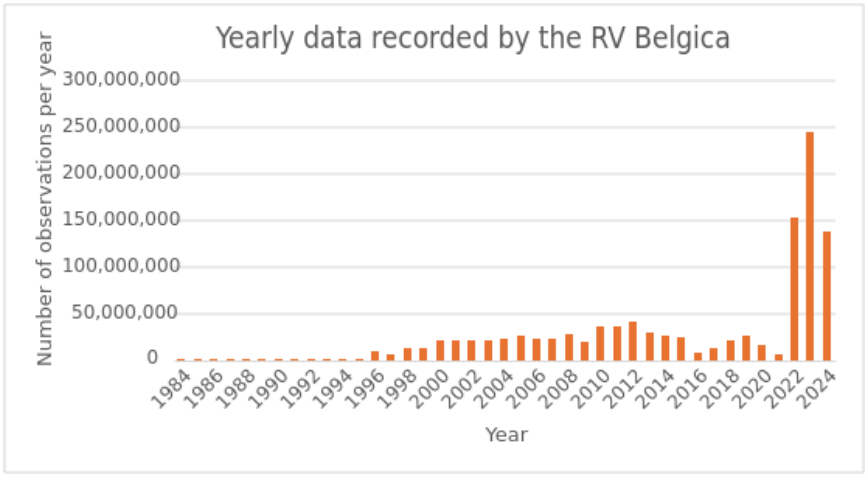


Figure 2: Number of observations recorded in the central database of the Research vessel (1984 - June 2024). From 1984 to 2021, the data were collected by the previous A962 RV Belgical. Since 2022, the data are collected by the new RV Belgica.

Those millions of observations represent only the en-route data, which are stored in the central vessel database. These observations can be grouped into four categories: meteorological, navigational, oceanographic and hydrographical data (Figure 3). In addition to this, large volumes of heavy data, such as (scientific) multi- and singlebeam echosounder readings, hydroacoustic current meter readings and sub-bottom profiler readings, are generated in substantial quantities.

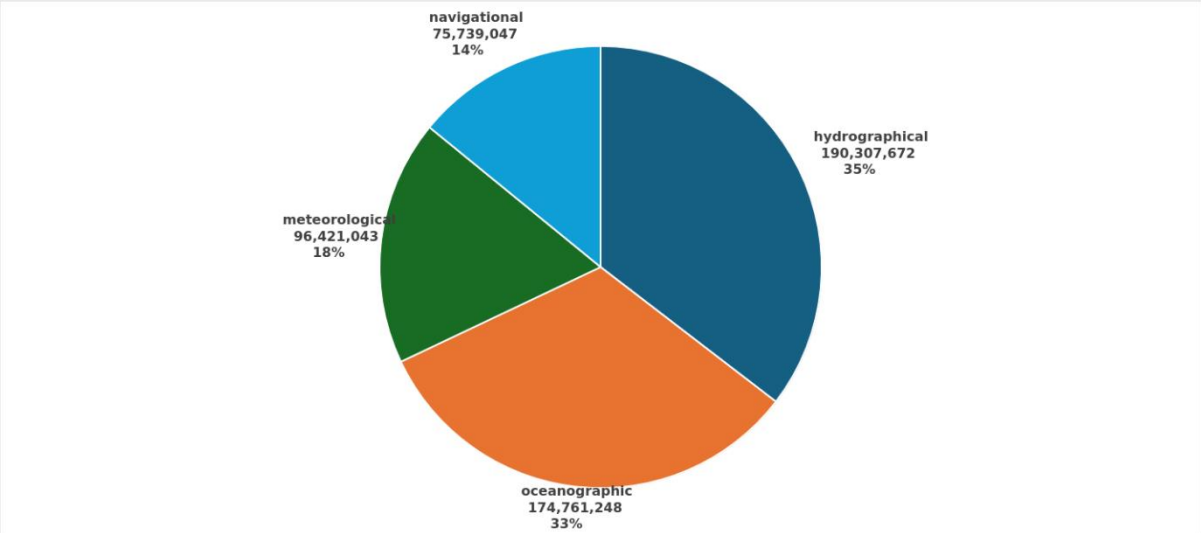


Figure 3: Thematic distribution of the data recorded by the RV Belgica. For each category, the total number of observations is shown (January 2022- June 2024) along with the percentage of the total number of observations.

Given these challenges, the ODANext project represents a pivotal opportunity to renew and enhance the entire data acquisition workflow associated with the RV Belgica. By addressing the modern challenges posed by increased data volumes, evolving IT standards, and the need for greater interoperability, the project aims to elevate the visibility of the Belgian research vessel on the global stage. Furthermore, the improvements in data management and dissemination processes will significantly ease the work of researchers, facilitating more efficient and widespread distribution of scientific findings in the coming decades.

2. STATE OF THE ART AND OBJECTIVES

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, the water, the seabed and beneath. The RV Belgica is a floating laboratory, composed of dozens of sensors and instruments, and a key Research Infrastructure of the Institute of Natural Sciences and the Federal portfolio. These continuous measurements (time series of deployments and 'en-route data') relate to meteorological, navigation, physical/chemical and operational parameters. They support (fundamental and policy-supportive) research, provide calibration for mathematical models, and provide long-term insights in the changing conditions of the North Sea.

Currently the RV Belgica stores its en-route data in an outdated database system, designed in 1996 (ODASIII, Oceanographic Data Acquisition System). The new vessel must adopt a state-of-the-art continuous data management system, which is the focus of this project.

Continuous data serves a wide range of users: researchers, the global marine research community, private companies, in-house dissemination tools and various Data infrastructures that RBINS contributes to (e.g. GOSUD, INSPIRE, ICOS, SeaDataNet,...). To meet the needs of these diverse users,

a fully automated sensor-to-client data flow has been developed for all bound sensors, providing rich, standardized and quality-controlled data in near real-time.

The main objective of this project is to create an automatic data and metadata workflow for the new RV Belgica through collaboration between the marine data acquisition (MSO, Ostend), data management (BMDC, Brussels) and scientific websites (SWAP, Brussels) teams of the Institute of Natural Sciences.

With the coming of the new vessel, a novel data processing system, taking advantage of leading-edge technology in data-processing (O&M, SWE, Quality control...) is needed to tackle the requirements of each user. O&M refers to the Observation and Measurements standard while Sensor Web Enablement (SWE) is a suite of standards to make sensor information and data accessible via the web. Various efforts already made the data *Accessible* (notably the OD Nature website), but never up to the optimal level of the FAIR (*Findable*, *Accessible*, *Interoperable*, *Reusable*, Wilkinson M. *et al.* 2016) best practice.

These developments require a project-based, integrated approach involving collaboration between data providers and the data acquisition and management teams. The FAIR principles provide a framework for best practices. When applied to sensors, they promote interoperability and reusability through early, complete, and transparent sensor description ('metadata'), such as placement and calibration, and by using agreed terminology (*i.e.* controlled vocabularies). The project aims to improve data governance and transparency of publicly funded cruises.

Objectives:

- Fully operational sensor-to-client data flow for all sensors in near real-time
- Rich, standardized data that can serve any (potential) client
- Metadata enrichment when necessary
- Optimized and secure data storage
- Integration into relevant open science data repositories
- Enhanced data governance by writing DMPs in collaboration with data providers.

This project is innovative because it will completely overhaul the data processing workflow of environmental data collected by the new Belgica’s equipment. It will significantly improve the current in-house environmental data processing and align with state-of-the-art data management practices. Elements such as sensor descriptions, enrichment with linked data to controlled vocabulary, standardized quality flags and OGC standards will be included at the appropriate stages in the data workflow to ensure sustainable acquisition, storage, and dissemination of the data now and in the future. We will provide free and uniform access (both human and machine-to-machine) to the data, which will increase the citability of BELSPO and RBINS data by facilitating the creation of DOIs for specific data subsets.

3. METHODOLOGY

The delivery of the new research vessel, though delayed, marks a significant milestone in our scientific endeavors. This highly sophisticated tool, brimming with advanced technologies, presents a unique set of challenges and opportunities for testing and understanding. Initially, the methodological principles guiding this project were rooted in theory. Our theoretical framework was based on the MoSCoW methodology, prioritizing tasks as Must-have, Should-have, Could-have, and Won't-have, ensuring clarity in essential deliverables.

However, the real-world implementation demanded a more dynamic approach. Adopting Agile principles, we embraced flexibility and iterative progress, allowing us to adapt swiftly to the complexities of the vessel and the unpredictable nature of marine research. Yet the MoSCoW method was crucial in identifying components that could potentially block future development, thereby setting clear work priorities. The work was planned as cascading packages, meaning each phase depended on the successful completion of the preceding one. Therefore, it was crucial to ensure that downstream components were not blocked by delays in the initial stages.

The FAIR principles, which stand for Findable, Accessible, Interoperable, and Reusable, provided the backbone for all our choices throughout this project. By ensuring data and resources adhere to these guidelines, we aimed to enhance the efficiency and utility of our research outputs. Each decision, from data management to tool selection, was guided by the need to make information easily discoverable, readily accessible, compatible across systems, and reusable for future research. This approach not only streamlined our workflow but also ensured that our findings would have lasting value and applicability within the broader scientific community.

In the initial proposal, we had to include a contingency plan to address potential unexpected delays or issues, anticipating the inherent uncertainties of such a project. This foresight proved invaluable as we encountered a series of significant challenges. The COVID-19 pandemic disrupted timelines and operations globally, which further delayed the delivery of our research vessel. Additionally, the vessel's on-board systems were more immature than anticipated, requiring extensive troubleshooting and adjustments. Compounding these issues, we also faced the departure of several key colleagues. The robust contingency plan allowed us to navigate these obstacles effectively, ensuring that despite the setbacks, we could continue to advance the project and adapt our strategies to meet the evolving circumstances.

Throughout the project, we organized several key activities aimed at improving the quality of our results. Continuous testing of the components was a fundamental part of our approach, ensuring that each element was thoroughly evaluated and refined. To further enhance our capabilities, we organized a training session with a specialist consultant. This session provided us with in-depth knowledge and practical skills, allowing our team to gain a better understanding of the different technological options and their consequences for the result. To stay aware of the latest developments in our field, we also attended various international conferences. These events offered valuable insights into modern technological advancements and best practices, which we could integrate into our project. Additionally, they provided networking opportunities, allowing us to connect with experts and innovators.

Our engagement with the follow-up committee was another crucial aspect of our project. The committee includes specialists in on-board software and data management, as well as authorities on international data dissemination. These experts provided ongoing oversight and feedback, guiding our efforts, and ensuring alignment with our objectives. This collaborative relationship ensured that all aspects of our project were externally monitored and optimized, significantly improving the quality of the result.

Our software choice philosophy was grounded in a firm commitment to utilizing open-source solutions exclusively. This decision was driven by the flexibility, transparency, and collaborative potential that open-source software offers. While this approach presented some challenges, such as dealing with implementation and maintenance work and a steep learning curve, these drawbacks were effectively counterbalanced by the significant benefits. The learning curve provided our team with valuable new skills, enhancing our technical proficiency and problem-solving abilities. Additionally, the supportive open-source community offered great support and exchange of ideas, enabling us to leverage collective expertise and resources. This philosophy not only fostered innovation and adaptability but also aligned with our principles of open collaboration and continuous improvement.

Our project involved several choices that sparked discussions among the team. Initially, we focused on writing technical specifications in terms of desired results rather than implementation details, acknowledging that we would need to adapt and learn as we progressed. This flexible approach allowed us to answer the significant challenges related to performance, structural costs, and maintenance. For instance, we initially planned to use Sensor Observation Service (SOS) standard, but it soon became apparent that SOS could not handle the volume of data generated by the new research vessel efficiently. After extensive discussions, testing and consultations with the follow-up committee and experts we met at conferences, we decided to switch to SensorThings API, a more modern alternative to SOS. SensorThings is INSPIRE-compliant as well. This decision enhanced our system's performance and better suited our project's needs, demonstrating the value of flexibility and expert input in our decision-making process.

4. SCIENTIFIC RESULTS AND RECOMMENDATIONS

The project primarily focused on establishing an efficient data management system to support scientific activities conducted on-board the RV Belgica. Rather than generating scientific results, the aim was to design and implement a robust data architecture that facilitates the collection, processing, dissemination, and metadata management of the data gathered during the scientific cruises. This comprehensive infrastructure will ultimately enhance the quality and reliability of scientific results by ensuring that researchers have seamless access to well-documented and easily accessible data. It will also improve the quality of the data published in national and international data repositories. The following sections of this report will detail the data architecture we have designed and implemented, illustrating how we achieved the desired outcomes through various interconnected systems and platforms.

Figure 4 illustrates the integration of the new architecture developed for the RV Belgica with the existing infrastructure. The main goal is to enhance the impact of the new data workflow by leveraging the dissemination pathways established in previous projects. By adhering to the same implementation philosophy, we reduce future maintenance costs by consolidating the tools used. Any updates to the downstream (meta)data components will benefit all upstream nodes. Additionally, using open-source software with a strong community ensures that we receive updates without risking expensive software licenses, vendor lock–in or the need for in-house full-time developers.

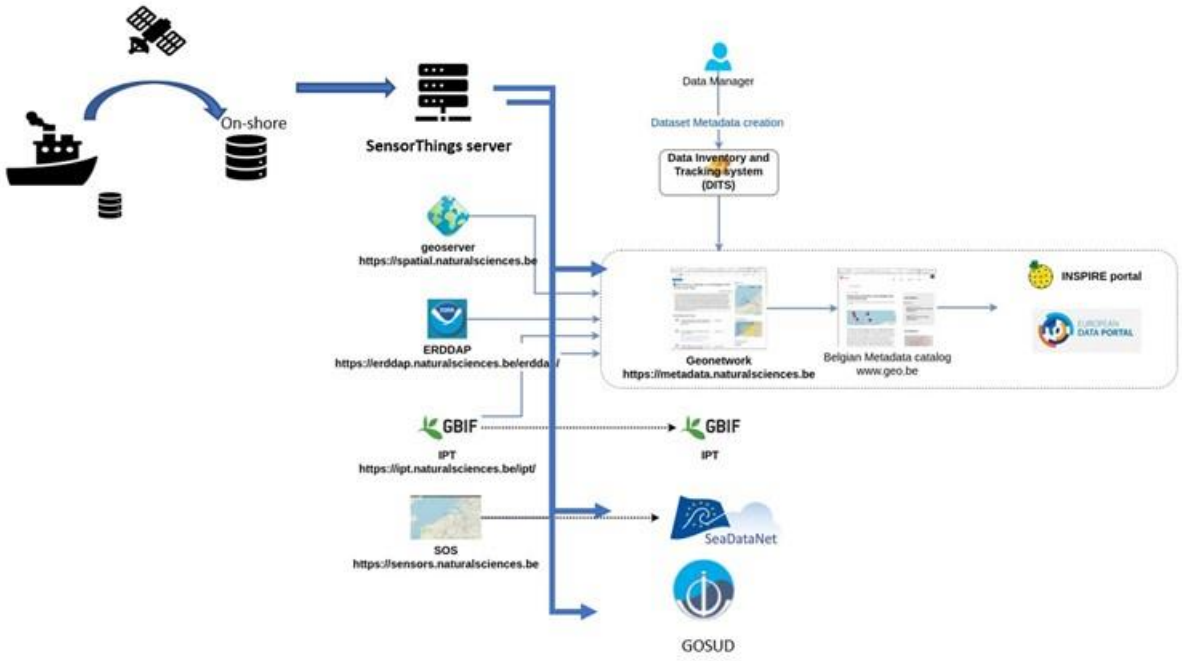


Figure 4: Presentation of the seamless integration of the new RV Belgica data workflow into the existing data and metadata infrastructure of the Belgian Marine Data Centre (BMDC). The blue arrows represent the new links implemented during the ODANext project.

The first step, which is quite complex, involves effectively managing the data on-board at the time of acquisition (hardware and software). This includes database design, optimization, and storage, along with the proper parametrization of various components. After acquisition, the data must be transferred to shore in a cost-effective manner and processed. This processing includes data

formatting, geo-referencing, metadata editing, and quality control with flagging before further dissemination. Due to the volume of data generated, optimization is essential at each step to prevent backlogs.

4.1. On-board MDM500 data acquisition software and database

The vessel was delivered with a Microsoft SQL database (licensed). The database design was solely focused on storing raw data, without normalization, geo-referencing, or quality checks. Our initial steps involved understanding the system's behaviour and performance and troubleshooting the various issues associated with such a new and complex vessel. This process presented a steep learning curve and took longer than anticipated.

Firstly, we examined all the drivers transferring data from the sensors to ensure that raw data were not mixed up in the database. We performed an exhaustive check of all 555 parameters, leading to numerous corrections. Next, we investigated the time gap between actual observations and their recording in the database, which required understanding and matching the sensor acquisition rate with the database writing capacity. We discovered that the system was overwhelmed by the data volume, so we upgraded all hard drives to faster SSDs and optimized the acquisition rate.

During the ODANext project, we explored alternatives to the built-in acquisition software, such as TechSAS 1 and TechSAS 2. After discussions with the main developers, it became clear that TechSAS 1 was becoming obsolete, and TechSAS 2 was not mature enough for installation on the RV Belgica. Consequently, we decided to continue using MDM500, despite its flaws, and to work with the developers to resolve as many issues as possible. We are now considering the open-source acquisition software RVDAS to determine if it could be a viable option for the RV Belgica, although this is beyond the scope of the current project.

An essential aspect of the on-board data acquisition system is the ability to visualize data from different sensors in real-time. This greatly enhances data quality, as faulty sensors are often easy to identify visually, whereas robustly and reliably automating such checks can be challenging. While obvious errors (such as no data recorded or significant offsets) can be automatically detected, nothing is as effective as an expert's eye. In collaboration with the vessel manufacturer, we decided to install open-source Grafana software (Figure 5), allowing users to select each sensor connected to the central database and visualize the data. We are continuing to improve the dashboards based on user feedback.

Figure 5: Example of the Grafana dashboard deployed on-board the RV Belgica for a real-time data visualization. This represents the depth measured by the ADCP sensor along with some navigational parameters.

4.2. Vessel-to-shore data transfer (automatic and manual) and On-shore database.

Different options for transferring data to shore include V-SAT, Iridium, and StarLink. When selecting a system, we must consider the volume of data to be transferred, the transmission frequency, the procedures available for handling faulty transfers (e.g., when cruising in areas without coverage or losing connection during transfer), and, of course, the cost. Based on our experience with the previous RV Belgica, where we used V-SAT, we decided to continue with the same system. The main advantages are its cost, reliability, and our familiarity with it. The relative disadvantage is the delay in data transfer. Initially, this was a concern due to the volume of data, but our experience has shown that V-SAT is more than sufficient for our needs. In addition to the automated data transfer, we make sure to download manually the cruise data each time the vessel comes back to the port of Zeebrugge.

Figure 6 presents a detailed schema of the data architecture setup for the RV Belgica. The on-board database is a Microsoft SQL database, a proprietary software (though there is a free version, it is limited to a small amount of data). To avoid licensing issues and remain consistent with our existing infrastructure, we created a replica of the MDM500 database in PostgreSQL for raw data import. We then developed a procedure to geo-reference each observation using the best available GPS information (out of three GPS sources) and format the data for import into the final PostgreSQL database hosted on our internal servers. The database schema follows the Open Geospatial Consortium SensorThings standard (OGC, 2016), specifically designed to handle sensor time-series data and facilitate machine-to-machine data dissemination. This standard includes all necessary fields to accurately describe sensors and observation metadata, such as units, observed properties (e.g., water temperature), and quality flags. Each database entry is described with controlled vocabularies

(e.g., NERC vocabularies) to support future automated data exchanges. We also imported all historical data from the previous RV Belgica (1984-2021) into the same database.

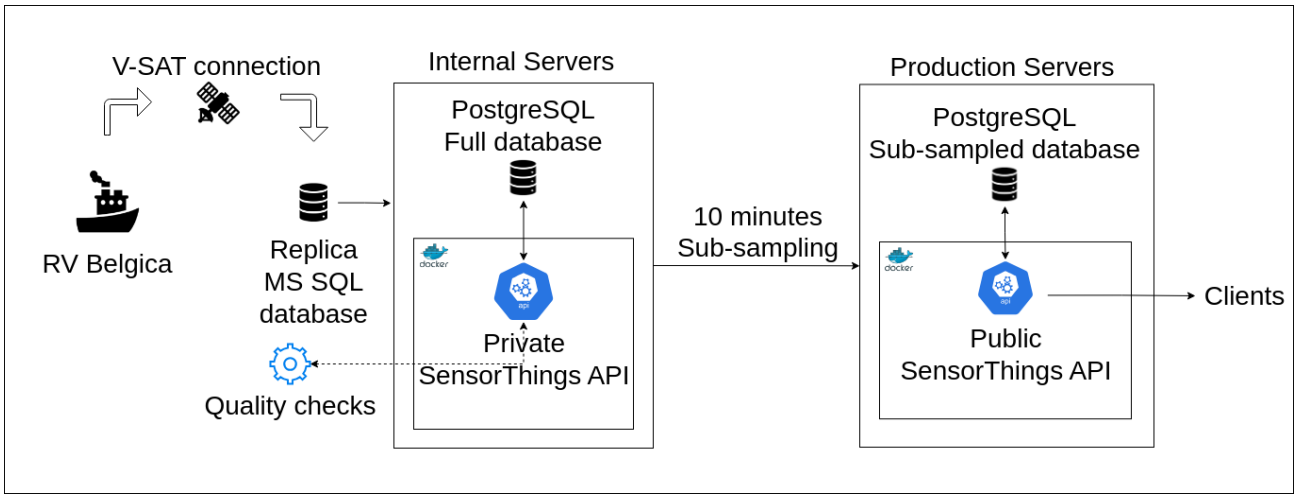


Figure 6: Detailed view on the data architecture designed and implemented to automatically manage, document and quality control the data in near real-time.

On top of the database, we deployed an open-source software developed by the Fraunhofer Institute called FROST SensorThings API. This software provides an API (Application Programming Interface) that exposes the data online following the OGC standard specification, allowing developers to access the data without any knowledge of the internal database structure. Sensor information can be fetched automatically alongside the data. The advantage of using an API is that if we need to change the database structure in the future, the interface to which the downstream software connects remains the same, ensuring no disruption in data dissemination. This API is only accessible on the internal network of RBINS for scientists and data managers.

As mentioned earlier, the volume of data recorded is significant, with most parameters recorded once every second. This frequency is necessary to meet all scientific requests but often results in data handling challenges. Most users need a lower frequency (e.g., one observation every 10 minutes). This is the case for the international repository GOSUD and users of the official RV Belgica website. We considered two approaches: creating duplicate time-series with lower frequency in the same database, which would create data duplication issues, or duplicating the infrastructure and importing a 10-minute sub-sampling of the database into another, structurally similar, database. We chose the latter approach, providing a lightweight database that is easy to share and performs well. The full database remains the reference where all checks are conducted. On top of this lightweight database, we deployed another SensorThings API exposed to everyone. By controlling what is exposed, we protect our infrastructure from automated external requests that could jeopardize stability.

For long-term data storage, we set up an automated backup procedure for the entire database on different servers and external tape. This ensures that even if the database is corrupted, we can recover the 40+ years of data. It is important to note that the on-board database system cannot be considered a safe backup location.

4.3. Optimization at all levels

Millions of observations are recorded daily while the vessel is at sea. This necessitates rigorous testing of each component of the data workflow to ensure it can handle the data stream efficiently, both now and in the future. During the ODANext project, we began working with the MDM500 database on board. Initially, the hardware was not capable of efficiently handling the 555 incoming time-series. After several test cruises, we observed that the system was constantly operating at full capacity, prompting us to replace the hardware with more efficient components. Additionally, we reduced the acquisition frequency for parameters with low variability, such as atmospheric temperature. This optimization decreased the number of daily observations from approximately 5 million to around 4 million, depending on which sensors are active.

For near real-time data transfer, the on-board software lacked a procedure to export and compress data. To address this, we developed a procedure that compresses and transmits data via V-SAT, reducing both bandwidth usage and cost.

The main optimization work occurred at the database level. One of the project's goals was to provide seamless access to both historical and new data, meaning all data should fit within the same system while maintaining good performance. Given the database's rapid growth, we needed a system that would remain efficient as data accumulated. Delays in the vessel's construction meant we couldn't test the system with new data initially. Fortunately, we had 40 years of historical data to test various database setups.
We started with a single database for historical data, which performed well. However, when the vessel began operations, the database grew rapidly, and we noticed a slight performance degradation. We considered moving everything to the cloud, which offers scalable infrastructure, but we rejected this option due to cost and our preference to keep the data close to the Institute, as it is the RV Belgica's most valuable asset. Instead, we capitalized on our existing database servers from other projects.
Our second approach was to create a second database only for the new RV Belgica. While this improved performance, it complicated seamless access to both sets of data and increased maintenance work.

The third and final approach that we kept was to install only one database for both research vessel with the TimescaleDB extension for PostgreSQL, which efficiently manages time-series data by partitioning the database. This ensures that performance remains robust even as the database grows.
In summary, simplicity is challenging to achieve. Complex systems should be simplified once they mature. As the project progressed, we added more elements, which we could eventually simplify thanks to our improved understanding of the available solutions.

Optimization should be goal-oriented. Creating complex and expensive systems to handle occasional requests is resource-intensive and unnecessary. User feedback was invaluable in this regard. Frequent machine-to-machine data requests (e.g., daily) typically require low-frequency data with quick access times (a few seconds). Full-frequency data requests mainly come from scientists needing data from their cruises, who typically request datasets once after their cruise. For these users, it is acceptable if the request takes a few minutes to process.

4.4. What about data quality?

The volume of data generated by modern, well-equipped research vessels is substantial. To assist data managers, automated tools for quality control are essential, allowing them to focus on in-depth analysis of suspicious data. However, automating quality control is challenging, as observations can only be accurately labelled as good or bad when viewed from multiple 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 through a series of pumps, tubes, and valves. To validate the data, we must ensure that the water flow is sufficient to renew the water being analysed by the sensors, ensuring its properties are not significantly altered (e.g., water temperature). All parameters depend on the water flow. Additionally, several parameters are computed based on others (e.g., salinity, sound velocity). Primary parameters must be validated first before checking the computed parameters.

Geo-Location Dependencies: Threshold checks are performed with caution and sufficiently wide boundaries, as water quality can change significantly when the vessel enters specific areas, such as estuaries or high-latitude seas. Thresholds should always be analysed in conjunction with gradients.

Spikes and Gradients: Parameters have a limited versatility over time and space. We check that the values are not changing too fast over time (gradient check). The accepted gradient values must be fine-tuned parameters by parameters with a trial-and-error approach. The work is still going-on to determine the best values. Spikes are easy to identify by computing the delta between each consecutive points.

Cross-Validation: Some parameters are measured multiple times by different sensors at various locations. This redundancy can be used to analyse suspect trends. If a trend is present in all sensors, it can be considered valid. If only one sensor identifies it, it should be flagged as suspect. This approach is not implemented in the automatic QC as it requires a visual check.

In conclusion, automated quality control is only part of the data management solution. We created an open-source package in Python to perform it (https://github.com/naturalsciences/qualityAssuranceTool). It helps exclude obvious errors and highlights suspicious data, drawing the data manager's attention to potentially interesting events. Figure 7 illustrates this process. The first graph shows suspicious water temperature data automatically flagged due to a steep gradient. The second graph shows that the water flow in the circuit was not disturbed and cannot explain the spike. The third graph shows the water temperature measured by a second independent sensor. Since the spike is also present in this data, it is validated, and the data can be flagged as "good" using SeaDataNet quality flags.

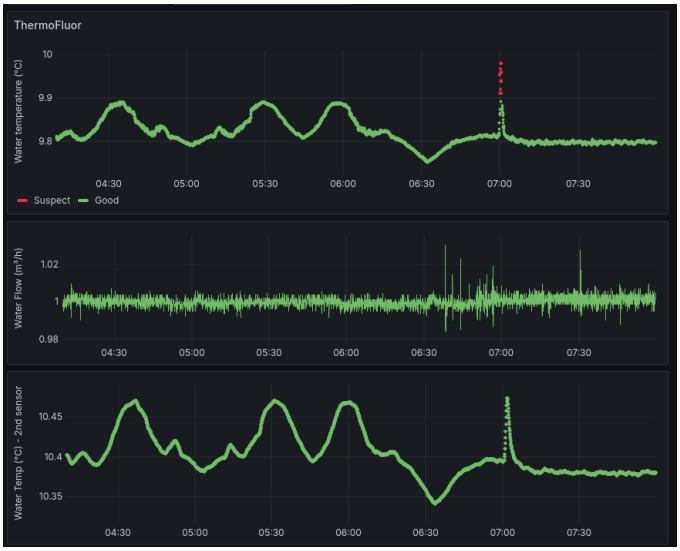


Figure 7: Visual analysis in Grafana of the data quality check performed automatically.

4.5. New Skills

During this project, our different teams involved acquired and honed several new skills that were crucial to its success. One of the most significant areas of growth was learning and mastering the technical aspects of the new research vessel. This involved understanding the vessel's complex systems, including sensor wiring and parametrization, data collection with its hardware and software, and communication technologies. Gaining proficiency in these areas required both hands-on experience and theoretical study, allowing us to operate the vessel efficiently and leverage its full capabilities for research purposes.

Another key skill developed during this project was working with big data. The sheer volume of data collected by the vessel's advanced sensors necessitated the use of efficient data management techniques. We learned how to optimize each component of the data pipeline, from collection and storage to analysis and visualization. This optimization was critical to ensuring that the data could be processed in a timely and accurate manner, ultimately leading to more informed decision-making and higher-quality research outcomes.

Crucially, our success in mastering these techniques was greatly supported by our international contacts. The follow-up committee, which included experts from various countries, provided invaluable guidance and feedback throughout the project. Additionally, colleagues we met at different conferences played a significant role in helping us refine our approaches and stay updated on the latest advancements in data management. These international collaborations not only enhanced our technical skills but also enriched our perspectives, allowing us to achieve our goals more effectively.

Implementing and testing new metadata and data standards was also a vital part of the project. As the volume and complexity of the data grew, it became clear that consistent and well-defined metadata was essential for maintaining data integrity and facilitating collaboration. We gained experience in creating and applying these standards, which involved understanding both the theoretical underpinnings and the practical challenges of data standardization. Testing these standards in real-world conditions further deepened our understanding, as we worked to ensure they were robust and adaptable to the specific needs of the project.

5. DISSEMINATION AND VALORISATION

The central objective of this project was to safeguard invaluable (historical) data while establishing a robust and durable infrastructure for data dissemination and valorisation. Recognizing the significance of preserving and making accessible decades of research, our efforts have been directed towards creating a system that not only ensures the longevity of this data but also facilitates its efficient distribution to a wide array of stakeholders. The successful dissemination of this data is necessary for advancing scientific research, supporting policy-making, and fostering collaboration across various disciplines.

In the project proposal, we identified the most important data users that we should be providing data to as a start, namely the RV Belgica website, the GOSUD, SeaDataNet and INSPIRE repositories. The following section describes briefly the results achieved for each end-user.

5.1. RV Belgica website

For the project, we completely rebuilt the website's data-related pages from the ground up, using a modern framework and leveraging the various functionalities of the SensorThings API. This approach allowed us to retain all the features of the old website while enhancing data accessibility by adding new filtering and search capabilities. The transition from the old to the new website is seamless for users, despite the new design making data more comprehensible and accessible, with features like on-the-fly time-series diagrams for each cruise variable and sub-sampling of relevant data for further download.

However, the primary achievement was the successful migration to the new API, which standardizes data access and download. This shift significantly improves the FAIRness of the data and increases the system's longevity by eliminating custom ad-hoc developments. The API also enables us to effortlessly distribute data from both the new and old RV Belgica, allowing users to easily select and analyze data from cruises dating back to 2004. For example,

Figure 8 displays data from the 2023/05 cruise, featuring the trajectory on a map and a couple of time-series (atmospheric temperature and dissolved O2 concentration).

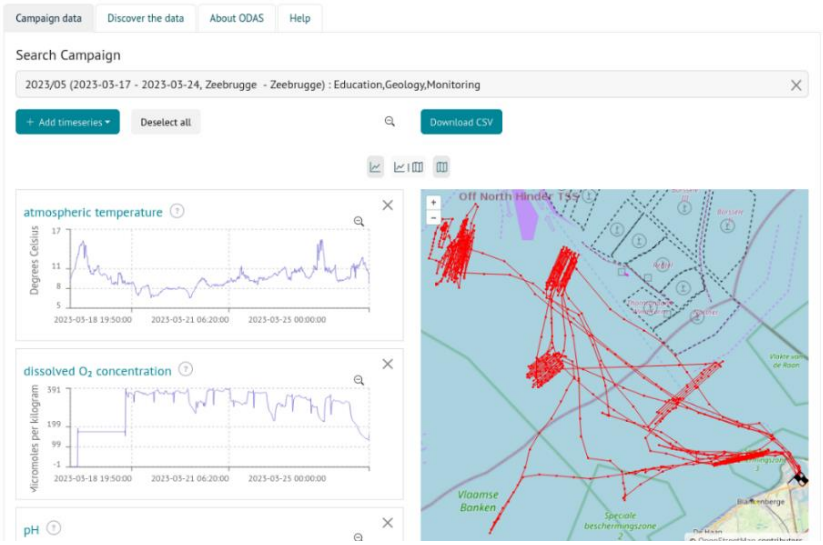


Figure 8: The new web interface to visualize, select and download the data for each campaign (2004-today).

The website can be accessed at: https://odnature.naturalsciences.be/odanext/en

5.2. GOSUD repository

The Global Ocean Surface Underway Data (GOSUD) Project, an Intergovernmental Oceanographic Commission (IOC) program, is designed as an end-to-end system for collecting and sharing data from ships as they traverse oceanic routes. Historically, we have contributed RV Belgica data to this program, but data transfer ceased a few years ago when the existing system became obsolete. As part of the ODANext project, we decided to resume data transfer, this time using standardized data formats and exchange protocols, specifically the OGC SensorThings standard. In collaboration with our colleagues at Ifremer, we developed a new procedure for fully automated, daily transfers of quality-controlled data. Figure 9 shows RV Belgica data displayed on the GOSUD website (https://www.gosud.org/Data-access/GOSUD-dashboard).

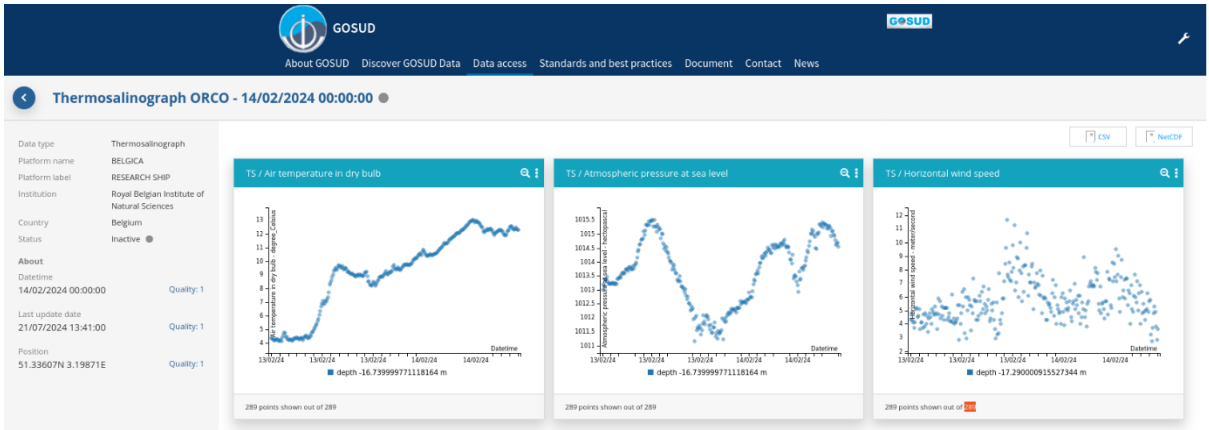


Figure 9: Web-interface of the GOSUD website. RV Belgica data are fetched daily from the new API automatically every day and shared on their European network.

5.3. SeaDataNet

SeaDataNet is a distributed Marine Data Infrastructure designed for managing large and diverse datasets derived from in situ observations of seas and oceans. RBINS has been a longstanding partner in the development of the SeaDataNet infrastructure and a frequent data contributor. As part of the ODANext project, our objective was to publish all relevant historical data and establish an efficient process for future data publication. We consolidated the data from both RV Belgica vessels into a single database, with all necessary metadata encoded as linked data using controlled vocabularies. This machine-to-machine description of data is a critical and often challenging step in data sharing. By incorporating these definitions and adhering to data standards from the outset, we ensure seamless dissemination across various data infrastructures with minimal additional effort. In SeaDataNet, data are published as datasets known as common data index (CDI, a metadata standard used to provide detailed descriptions of marine datasets). We have published 659 new thermo-salinometry datasets from the 'old' A962 RV Belgica and established the pathway for publishing data from the new RV Belgica.

5.4. INSPIRE

The Infrastructure for Spatial Information in the European Community (INSPIRE) Directive (2007/2/EC) aims to establish a European Union spatial data infrastructure to support EU environmental policies and any activities that may impact the environment. The Directive imposes requirements on public bodies that produce, receive, manage, or update spatial datasets covering all land and marine areas under the jurisdiction of member states, ensuring the creation of a cohesive EU-related spatial data infrastructure. The INSPIRE Directive was further strengthened by the adoption of the third PSI Directive 2019/1024 (also known as the Open Data Directive), which enhances data exchange capabilities.

The data collected by RV Belgica falls under the scope of both Directives. In a previous project, infrastructure for INSPIRE-compliant metadata publication was developed and is being reused in the current project. As a result, metadata describing this data is already available on the RBINS metadata catalog (https://metadata.naturalsciences.be), the national INSPIRE node managed by NGI (http://www.geo.be), and the European INSPIRE metadata portal (https://inspire-geoportal.ec.europa.eu/srv/eng/catalog.search#/home). The data is available for reuse under the conditions of the Creative Commons Public Domain Dedication (CC0) and is accessible via a standardized API, fully compliant with the Open Data Directive specifications.

5.5. UGent Master Thesis – Titus Top

As a final example of valorisation, we wanted to emphasize the work of Titus Top for its master thesis at UGent entitled "*Identifying trends of sea surface temperatures in the Belgian part of the North Sea due to climate change*".

By using the data collected by the previous RV Belgica and made accessible thanks to the work performed in the ODANext project, T. Top could perform a deep statistical analysis of the sea surface temperature evolution in the past four decades. The method and detailed results can be found in its thesis, but the most eloquent figure is shown here.

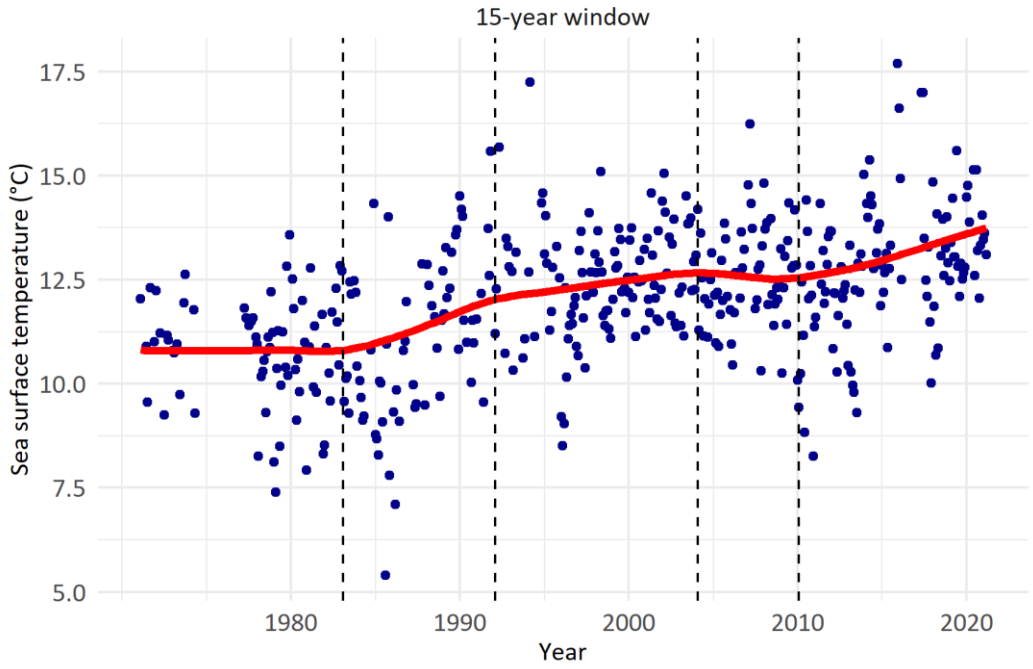


Figure 10: Long term sea surface temperature evolution in the Belgian part of the north Sea. (Top T., 2024)

6. PUBLICATIONS

de Ville de Goyet, N., & Vandenberghe, T. (2024). Deliverable 2.1.2 - Contribution on high-level user functionality of all on-board and on-shore systems to specification. (Version 1). Zenodo. https://doi.org/10.5281/zenodo.13373567

de Ville de Goyet, N., & Vandenberghe, T. (2024). Deliverable 2.1.3 – Investigating interoperability of O&M with common dissemination formats (INSPIRE, SeaDataNet ODV/NetCDF and GOSUD). (Version 1). Zenodo. https://doi.org/10.5281/zenodo.13373913

de Ville de Goyet, N. (2024). Deliverable 2.2.1 – Analysis of MDM 500, TechSAS1 and/or 2. (Version 1). Zenodo. https://doi.org/10.5281/zenodo.13373962

de Ville de Goyet, N., & Vandenberghe, T. (2024). Deliverable 2.4.2 – ETL specifications. (Version 2). Zenodo. https://doi.org/10.5281/zenodo.13373982

Vandenberghe, T., & de Ville de Goyet, N. (2024). Deliverable 2.5.2 – Sensor ML editor selection justification including proof-of-concept installation. (Version 1). Zenodo. https://doi.org/10.5281/zenodo.13374014

Vandenberghe, T. (2024). Deliverable 2.6.1 – Historical data handling specifications. (Version 1). Zenodo. https://doi.org/10.5281/zenodo.13374031

de Ville de Goyet, N., & Vandenberghe, T. (2021). Deliverable 3.1.1 – Tender call description: https://ted.europa.eu/fr/notice/-/detail/442413-2021

Youdjou N., Vandenberghe, T., & de Ville de Goyet, N. (2023). Deliverable 5.1.1 – Displaying the new data on the RV Belgica website. https://odnature.naturalsciences.be/belgica/en/

de Ville de Goyet, N. (2023). Deliverable 5.2.1 – Implementation of a SOS webservice communicating with the new database. https://sensors.naturalsciences.be/sta/

de Ville de Goyet, N. (2023). Deliverable 5.3.1 – Publication to GOSUD. https://www.gosud.org/Data-access/GOSUD-dashboard

de Ville de Goyet, N. (2023). Deliverable 5.3.2 – Publication to SeaDataNet. https://cdi.seadatanet.org/search/welcome.php?query=2967&query_code={F84982D0-84BD-4FC5-8371-E0C883156682}

de Ville de Goyet, N., & Vandenberghe, T. (2024). Deliverable 7.1 – Data provider and client survey. (Version 1). Zenodo. https://doi.org/10.5281/zenodo.13374047

de Ville de Goyet N., Van den Steen N. & Vandenberghe T (2024). Real-Time Data Transfer and Management for the RV Belgica Using FROST OGC SensorThings API. In Simoncelli, S., Vernet, M., & Coatanoan, C. (2024). International Conference on Marine Data and Information Systems - Proceedings Volume. Miscellanea INGV, 80. 75-76 https://doi.org/10.13127/MISC/80

7. ACKNOWLEDGEMENTS

ODANext is funded under the BRAIN-be.2.0 program of the Federal Science Policy (BELSPO) with the grant number B2/202/P2/ODANext.

We would like to acknowledge the support of all those who, in various ways, have contributed to the successful completion of this project. We also deeply appreciate the participation and feedback of colleagues in the follow-up committee of ODANext. Their guidance and suggestions helped us significantly improve our research.

The follow-up committee included:

André Cattrijsse, VLIZ, Belgium.
Louise Darroch, BODC, UK.
Koen Degrendele, FPS Economy, Belgium.
Nathalie Delattre, IGN, Belgium.
Michael Fettweis, RBINS, Belgium.
Michèle Fichaut, IFREMER, France.
Hans Hillewaert, ILVO, Belgium.
Ruth Lagring, RBINS, Belgium.
Adam Leadbetter, Marine Institute, Ireland.
Peter Thijsse, Maris, The Netherlands.
Vera Van Lancker, RBINS, Belgium.

REFERENCES

Top, T. (2024). Identifying trends of sea surface temperatures in the Belgian part of the North Sea due to climate change. Master of Science in de bio-ingenieurswetenschappen: land, water en klimaat. https://lib.ugent.be/catalog/rug01:003200765

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018. https://doi.org/10.1038/sdata.2016.18

Open Geospatial Consortium. (2016). *OGC SensorThings API Part 1: Sensing (OGC 15-078r6)*. https://doi.org/10.25607/OBP-1744