SOCIB Data Infrastructure

Juan Gabriel Fernández, SOCIB (Spain), jfernandez@socib.es
Paz Rotllán, SOCIB (Spain), protllan@socib.es
Miguel Charcos, SOCIB (Spain), mcharcos@socib.es
Biel Frontera, SOCIB (Spain), bfrontera@socib.es
Xisco Notario, SOCIB (Spain), xnotario@socib.es
Inmaculada Ruíz, SOCIB (Spain), iruiz@socib.es
Miquel Àngel Rújula, SOCIB (Spain), mrujula@socib.es
Miquel Gomila, SOCIB (Spain), mgomila@socib.es
Sonia Gómara, SOCIB (Spain), sgomara@socib.es
Joaquín Tintoré, SOCIB and IMEDEA (Spain), jtintore@socib.es

Background

The Balearic Islands Coastal Ocean Observing and Forecasting System (SOCIB), is a Marine Research Infrastructure, a multi-platform and integrated ocean observing and forecasting system that provides streams of data, added value products, and forecasting services from the coast to the open ocean (Tintoré et al., 2013). The SOCIB Data Center is in charge of managing the SOCIB data assets and their data life cycle and supporting the SOCIB products and services strategy. The SOCIB Data Management Program is the framework where the Data Center aims to meet all these requirements.

A new Data Quality Strategy

SOCIB Data Quality Strategy is being re-addressed to better meet the new requirements from the coastal ocean community and to keep up-to-date the IT systems that power SOCIB. This new Data Quality Strategy is driven by the improvement in SOCIB of both (1) data assets FAIRness and (2) data quality processes and procedures. To guarantee that SOCIB benefits from the new Data Quality Strategy, the SOCIB Data Management Program is being updated accordingly. Although not fully implemented, the upgraded Data Management Program is currently offering new tools and processes that improve the quality of the SOCIB data assets.

An improved Data Management Program

Updates on the Data Management Program pivot around two main action lines. The first one includes 3 processes: Quality Assurance, Quality Control and Quality Assessment. The second one requires upgrading the IT architecture to assure data quality. These 2 actions are shown in the figure 1.
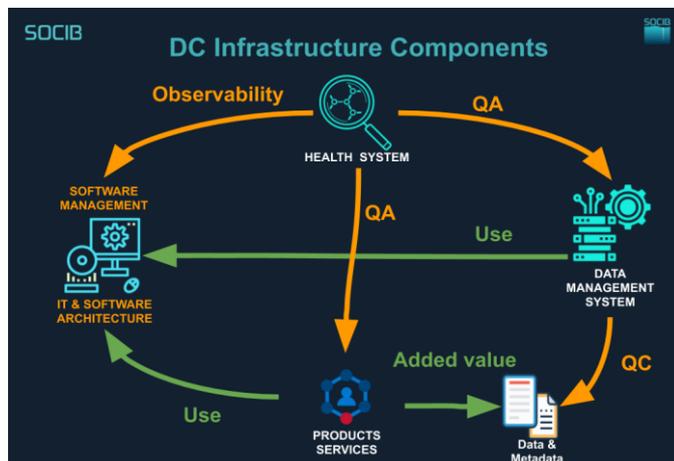

Figure 1: Data Center infrastructure components (orange represents the main improvement actions within the Data Management Program)

Within the SOCIB 2020-2021 Data Management Program update, the following action lines can be highlighted:

- Development of a *beta* Data Health Dashboard to facilitate Data Quality Assessment. The data quality assessment system integrates this dashboard, an autonomous Data Quality Checker and a custom Incidences Management System.
- Adoption of DevOps methodologies and tools to ensure both software quality and reliability of the SOCIB e-infrastructure. This also includes a migration to a container-based architecture (ie. Docker and Kubernetes).
- Improvement of the e-infrastructure observability (logs, metrics, performance, alerts) by adopting state of the art technologies.
- Improvement of the data life cycle observability: real time QC metrics, delay of arrival alerts and new data stewardship processes, among others.

In this presentation the above items will be introduced along with their role in the quality assurance of the SOCIB data assets. In addition 2 more action lines will be also introduced, showcasing downstream capabilities directly benefiting from the underlying data infrastructure:

- Release of an improved Data Catalog portal (http://apps.socib.es/data-catalog/), which includes embedded metadata in JSON-LD format (compliant with the DataCite metadata scheme), allowing Google Dataset to interoperate and discover the SOCIB data products (ie. the "I" of FAIR).
- Implementation of a linked data resource catalog to be integrated in the new SOCIB corporate website. There are also future plans to implement a linked data node.

Supporting processes

The new 2020 SOCIB Data Management Program will benefit from two strategic actions: (1) the implementation of a Quality Management System following the IODE's Quality Management Framework, to become an accredited Associate Data Unit of the World Data System (WDS), and (2) the CoreTrustSeal certification for the SOCIB data repository (with the support of the FAIRsFAIR project). Both certification processes are quite related to each other, since they focus on documenting quality processes of data management interconnected areas. Achieving the requirements of both certifications provides a better understanding of the complexity of the system and the opportunity to improve the processes, documentation and team management.

Conclusions

It is more and more evident today that existing ocean data systems have to meet new quality assurance requirements from the community they support. In addition, the evolution of both (1) existing technologies and methodologies, and (2) the emergence of new ones, create new core or internal requirements. In this scenario, the SOCIB Data Infrastructure, after a process assessment, is being upgraded with new components, processes and systems with the main goal of staying upfront the state of the art of leading international data infrastructures.

Tintore, J. et al. (2013), The Balearic Islands Coastal Ocean Observing and Forecasting System Responding to Science, Technology and Society Needs, Marine Technology Society Journal, 47 (1), doi: 10.4031/MTSJ.47.1.10