ID41 A DATA MANAGEMENT E-INFRASTRUCTURE FOR THE OBSEA CABLED OBSERVATORY

ENOC MARTÍNEZ[26], ALBERT GARCÍA-BENADÍ[155], DANIEL M. TOMA[24], MATÍAS CARANDELL[4], MARC NOGUERAS[17], JOQUÍN DEL RÍO[19]

ABSTRACT

Underwater cabled observatories are a key assets to monitor the oceans, providing high-resolution multi-parametric data from a wide variety of sensor systems. Their outstanding observational capabilities lead to significant amounts of data that need to be properly acquired, archived, curated and distributed. This paper presents the OBSEA e-Infrastructure, a modular data infrastructure to manage and distribute data from the OBSEA underwater observatory in a Findable, Accessible, Interoperable and Re-usable manner.

Keywords - Data Management, FAIR principles, Open Geospatial Consortium, SensorThings API, Cabled Observatories

INTRODUCTION

Underwater observatories are becoming a mature technology in the past decades, providing massive amounts of multiparametric data over large periods of time. Their ability to measure biogeochemical and physical parameters make them a valuable asset to understand the oceans. However, the management of such volumes of information proves a real challenge for small and medium-sized institutions. In addition to sensor data, activities such as maintenance, calibrations and deployment operations also need to be properly documented and archived. Applying the FAIR guidelines to this multidisciplinary and dynamic (meta)data is therefore a complex task [1].

An example of such installations is OBSEA, a cabled underwater observatory, located off-the-coast of Vilanova i la Geltrú (Spain) [2]. Since its deployment in 2009 it has been continuously acquiring heterogeneous environmental data such conductivity, temperature, depth, currents, waves, video, underwater sound, seismic activity physicochemical data among others. With more than 10 years of archived data, a modular and scalable e-infrastructure was required to manage bohth historical and real-time data. Furthermore, since OBSEA is part of several European initiatives such as EMSO, the integration of OBSEA's data services with further research infrastructures and data aggregators is required.

Different software tools and strategies have been proposed to address many of the data management, reducing the need of developing ad-hoc solutions. However, most of these tools address partially the requirements of data infrastructures such as data storage, access, visualization, alarming, etc. Therefore, data managers and sensor operators rely on several systems for their daily operations. This work proposes an e-infrastructure that leverages existing open-source and communityaccepted tools into a coherent and organized (meta)data workflow.

OBSEA E-INFRASTRUCTURE

This e-infrastructure, depicted in Fig. 1, aims to reconcile two different aspects: support the daily activities of sensor operators and provide access to high-quality (meta)data. The former includes the management of operations (sensor calibrations, deployments, maintenance operations, etc.) while the latter provides (meta) data access following the FAIR guidelines to both human and machines. Since all the data acquired are usually sent to shore in real-time, all the elements in the data pipeline (acquisition, quality control, processing and storing) are automated. Human intervention is only needed to register new equipment or operations (deployment, calibration, etc.).

When sensor data comes into the e-infrastructure it is processed by a set of data acquisition scripts. These are in parse the data and performs some preliminary checks. Seeveral tools are used, such as the SWE Bridge universal driver. The acquired data is then passed to the Real-Time Quality Control System (real-time QC in Figure 1), which adds quality information based on the QARTOD guidelines [3].

Once flagged, data is sent to OBSEA's e-infrastructure central comonent: the FROST Server (Fraunhofer Open-Source SensorThings Server), an implementation of the OGC SensorThings API (STA)[4], [5]. This standard from the Open Geospatial Consortium (OGC) defines a geospatial-enabled and unified way to interconnect the Internet of Things (IoT) devices, data and applications over the Web. Its flexible and powerful data model provides an ideal framework to encode all (meta)data components from complex sensor systems.

The e-infrastrucutre takes advantage of the expandable nature of the STA data model, encoding both contextual and operational metadata. Contextual metadata includes descriptive information about sensors and measuring stations: e.g. serial number, deployment information, contact person or sensor history. All this metadata is semantically enhanced by the use of controlled vocabularies from the NERC Vocabulary Service and provides machine-understandable context and meaning [6].

In addition to contextual metadata, operational metadata includes all the information required to automatize the workflow, such as quality control thresholds, instructions to integrate variables into datasets, scheduled tasks to be performed to a data streams (e.g. averaging), etc. Therefore, both (meta)data and configuration parameters are accessible through the SensorThings API, enhancing the traceability and data provenance.

The data ingested at the SensorThings API is periodically exported into ERDDAP and CKAN services [7], [8]. ERDDAP is a de-facto standard in the ocean observing community which provides standardized access to subsets of data in multiple formats. This service is used to ingest OBSEA's data into European research infrastructures such as EMSO and as a gateway for aggregators (MonGOOS, EMODnet, Copernicus, etc.).

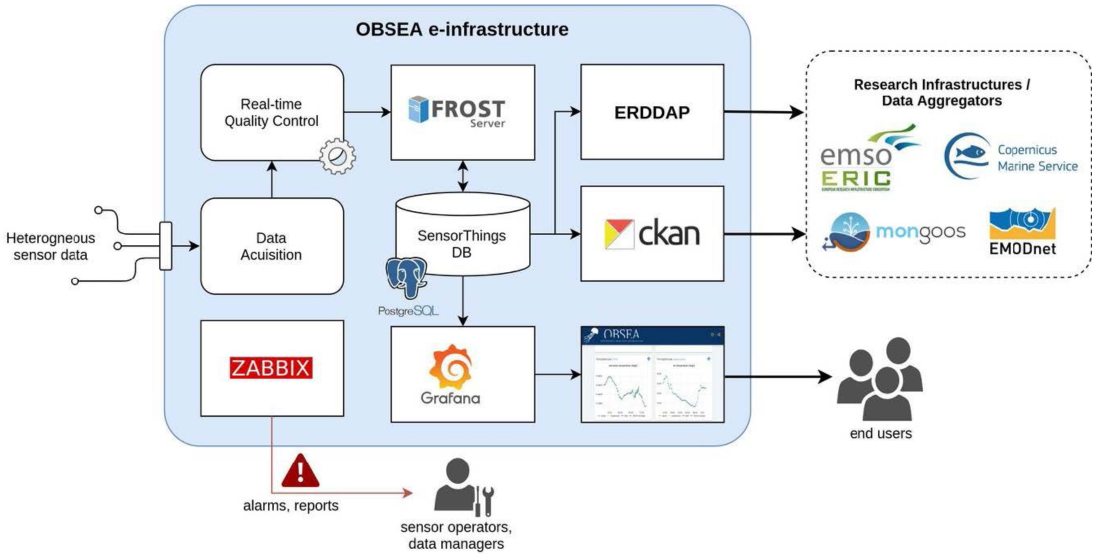
OBSEA e-infrastructure
Real-time Quality Control
FROST Server
ERDDAP
Research Infrastructures / Data Aggregators
emso ERIC
Copernicus Marine Service
Heterogeneous sensor data
Data Acuisition
SensorThings DB
PostgreSQL
ckan
mongoos
EMODnet
ZABBIX
Grafana
OBSEA
end users
alarms, reports
sensor operators, data managers

Fig 1. OBSEA e-Infrastructure dataflow and services. The incoming data from the sensors is parsed, quality-controlled in real-time and stored into a PostgresQL database through a SensorThings API. Data is periodically exported to data services such as ERDDAP and CKAN, which are used as a gateway to data aggregators. For visualization purposes data a grafana service is deployed and connected with OBSEA web page. A zabbix service monitors the e-Infrastructure health and the data dataflow.

CKAN (comprehensive knowledge archive network) is an open-source data management system that includes a powerful data catalog system. Within OBSEA's e-infrastructure, CKAN is the final storage of the produced datasets, where DOIs are assigned and maintained. In addition to CKAN's web interface, it also provides a powerful API, granting access to its data and metadata to both humans and machines.

For visualization purposes, OBSEA's data is also connected to a Grafana service, where dashboards are generated for both end-users and operators to internally asses the performance of the system [9]. Finally, the whole system health is monitored by a Zabbix alarming system, sending alarms and reports to the operators to quickly address any unforeseen situations [10].

CONCLUSIONS

In summary, the proposed e-infrastructure leverages existing software tools in order to achieve a complete FAIR data management system. Its data is findable since each dataset in CKAN has an assigned DOI and integrated to other research infrastructures such as MonGOOS, EMODnet, Copernicus and EMSO. Its variety of data access interfaces (OGC SensorThing, ERDDAP and CKAN) make (meta)data accessible to both human and machines. It is interoperable, since it is semantically enhanced by controlled vocabularies using common formats. Finally, it is reusable due to the open licenses used, (cc-by) and thanks to the extensive metadata on data provenance provided.

ACKNOLEDGEMENTS

This work has been financially supported by the European Commission's Horizon 2020 programme under the Jerico-S3 project (grant agreement 871153) and the Geo-INQUIRE project (grant agreement 101058518). This work used the EGI infrastructure with the dedicated support of EGI-CESGA-STACK. EM and MC acknowledge the financial support from the Ministerio de Ciencia e Innovación under the requalification of the Spanish university system program (Margarita Salas grant).

REFERENCES

[1] T. Tanhua et al., "Ocean FAIR Data Services," Front. Mar. Sci., vol. 6, p. 440, Aug. 2019, doi: 10.3389/fmars.2019.00440. [2] J. Del-Rio et al., "Obsea: A Decadal Balance for a Cabled Observatory Deployment," IEEE Access, vol. 8, pp. 33163–33177, 2020, doi: 10.1109/ACCESS.2020.2973771.

[3] M. Bushnell et al., "Quality Assurance of Oceanographic Observations: Standards and Guidance Adopted by an International Partnership," Front. Mar. Sci., vol. 6, no. November, pp. 1–12, Nov. 2019, doi: 10.3389/fmars.2019.00706.

[4] S. Liang, C.-Y. Huang, and T. Khalafbeigi, "OGC SensorThings API Part 1: Sensing," 2016. [Online]. Available: http://www.opengis.net/doc/is/sensorthings/1.0.

[5] H. van der Schaaf and M. Jacoby, "FROST Server." Fraunhofer Institut IOSB, 2016, Accessed: Apr. 12, 2023. [Online]. Available:https://github.com/FraunhoferIOSB/FROST-Server.

[6] Natural Environment Research Council, "The NERC vocabulary server: version 2.0," 2016. http://vocab.nerc.ac.uk/ (accessed Apr. 12, 2022).

[7] R. A. Simons, "ERDDAP." NOAA/NMFS/SWFSC/ERD., Monterey, CA, USA, 2019, Accessed: Jul. 18, 2020. [Online]. Available: ttps://coastwatch.pfeg.noaa.gov/erddap.

[8] Open Knowledge Foundation, "CKAN." https://ckan.org/ (accessed Apr. 12, 2023). [9] Grafana Labs, "Grafana." https://grafana.com.

[10] Zabbix LLC, "Zabbix." https://www.zabbix.com (accessed Apr. 12, 2023).