EGU2020-21634
https://doi.org/10.5194/egusphere-egu2020-21634
EGU General Assembly 2020

The data cube system to EO datasets: the DCS4COP project

João Felipe Cardoso dos Santos, Dimitry Van Der Zande, and Nabil Youdjou
Royal Belgian Institute of Natural Sciences, Brussels, Belgium (jcardoso@naturalsciences.be)

Earth Observation (EO) data availability is drastically increasing thanks to the Copernicus Sentinel missions. In 2014 Sentinel data volumes were approximately 200 TB (one operational mission) while in 2019 these volumes rose to 12 PB (nine operational missions) and will increase further with the planned launch of new Sentinel satellites. Dealing with this big data evolution has become an additional challenge in the development of downstream services next to algorithm development, product quality control, and data dissemination techniques.

The H2020 project 'Data Cube Service for Copernicus (DCS4COP)' addresses the downstream challenges of big data integrating Copernicus services in a data cube system. A data cube is typically a four-dimensions object, with a parameter dimension and three shared dimensions (time, latitude, longitude). The traditional geographical map data is transformed into a data cube based with user-defined spatial and temporal resolutions using tools such as mathematical operations, sub-setting, resampling, or gap filling to obtain a set of consistent parameters.

This work describes how different EO datasets are integrated in a data cube system to monitor the water quality in the Belgian Continental Shelf (BCS) for a period from 2017 to 2019. The EO data sources are divided in four groups: 1) high resolution data with low temporal coverage (i.e. Sentinel-2), 2) medium resolution data with daily coverage (i.e. Sentinel-3), 3) low resolution geostationary data with high coverage frequency (i.e. MSG-SEVIRI), and 4) merged EO data with different spatial and temporal information acquired from CMEMS. Each EO dataset from group 1 to 3 has its own thematic processor that is responsible for the acquisition of Level 1 data, the application of atmospheric corrections and a first quality control (QC) resulting in a Level 2 quality-controlled remote sensing reflectance (Rrs) product. The Level 2 Rrs is the main product used to generate other ocean colour products such as chlorophyll-a and suspended particulate matter. Each product generated from the Rrs passed by a second QC related to its characteristic and improvements (when applied) and organized in a common data format and structure to facilitate the direct integration into a product and sensor specific. At the end of the process, these products are defined as quality-controlled analysis ready data (ARD) and are ingested in the data cube system enabling fast and easy access to these big data volumes of multi-scale water quality products for further analysis (i.e. downstream service). The data cube system grants a fast and easy straightforward access converting netCDF data to Zarr and placing it on the server. In Zarr datasets, the object is divided into chunks and compressed while the metadata are stored in light weight .json files. Zarr works well on both local filesystems and cloud-based object stores which makes it possible to use through a variety of tools such as an interactive data viewer or jupyter

notebooks.