Technical documentation

Granularity Perspectives Document

Version 0.0, 2023-11-29: Draft

Versions

Version	Date	Comment	Responsible
0.0	2023-01-15	Draft	Øystein Godøy

Table of Contents

1. Introduction

1.1. Background

The intention of the GCW Data Portal is to be the entry point to datasets describing the cryosphere and form the information basis for the assessment activities of the Global Cryosphere Watch. It offers a web interface that contains information about datasets through discovery metadata provided by the data providers (or host data centre). These discovery metadata are harvested on a regular basis from data centres actually managing the data on behalf of the owners/providers of the data.

The GCW Portal utilises interoperability interfaces to metadata and data in order to provide a unified view on the datasets that are relevant for GCW activities. It is also the interface for GCW metadata to WMO Information System (WIS) and WMO Integrated Global Observing System (WIGOS)[1].

The GCW Portal will facilitate real time access to data through Internet and WMO operational data exchange[2] as requested by the user community. Consequently the GCW Portal will act as broker where necessary, between the data providers and the WMO Information System. This require a certain level of interoperability at the data level in addition to at the metadata level.

As the GCW community consists both of the normal WMO community and independent research organisations, GCW has focused on development of software components that allows GCW contributing stations to publish their data in a standards compliant way with a relatively limited time investment. This system is based on various technologies and software that have been developed over many years by some contributors to the GCW effort.

In order to improve interaction internally within GCW and ensure a proper information flow to the relevant GCW activities, some common perspectives on data granularity and how this is conveyed through discovery metadata is needed.

This document gives some perspectives on dataset granularity from the perspective of establishing integration services for GCW.

1.2. Applicable documents

RD-1 GCW Architecture Design Document (Not public available)

RD-2 GCW Interoperability Guidelines

2. Aspects on dataset granularity

2.1. Background

Granularity is a constant challenge in distributed data management. Quite often the discovery level information provided is only useful for human interpretation and not for machine action which is necessary to establish aggregated datasets and to let computers do the data preparation while humans can focus on interpretation and analysis.

Furthermore, dataset granularity if often defined by the data provider (i.e. the data collector/generator) which often has a different perspective on data than a potential data consumer. An example on this is e.g. profiles of temperature and salinity data from research cruises. These data are often published as a cruise dataset while a data consumer working with e.g. a numerical model is not necessarily interested in the cruise itself, but a combination of profiles from a large number of cruises in a specific region and time period. By publishing data based on research cruises, the work load on the data consumer is increased without any benefit. The visibility of the cruise could be established using parent/child relations between datasets or even by just tagging datasets with cruise reference. The is just an example from ocean data publishing, but similar examples could be identified from other domains.

NOTE A critical aspect of data publication is to acknowledge that data documentation and publication are done for anonymous data consumers we don't know and not for our own purpose only.

2.2. Granularity in practise

On a general basis the highest possible granularity (i.e. most detailed separation of data into datasets) is beneficial for machine action and reuse in e.g. numerical models, satellite calibration and validation activities etc. This implies not collecting many stations into the same dataset/file, but keeping them in separate datasets. These datasets can be combined into networks, field activity, research cruises etc, using parent/child relations on datasets or just tagging them with specific keywords.

Another aspect that should be avoided, is combining data at different time scales in the same dataset. E.g. if a station has some sensors observing at minute level and others at hourly basis, these should be separated in 2 datasets. Combining these in the same dataset is of course possible, but complicates downstream machine handling of multiple datasets as the complexity increases and adds too many degrees of freedom for the software aggregating data.

The same is also valid for feature types (e.g. Time series and time series of profiles) should not be mixed. Even though e.g. weather stations often have a mix of instruments with different characteristics where some instruments observe scalar values at a specific height (e.g. air temperature), others can observe profiles (e.g. permafrost). To avoid too many degrees of freedom for the software that is combining data and simplifying combinations these should be published as separate datasets.

NOTE The ambition by restricting how data are published is to simplify the work to pull different datasets into aggregated datasets in space and time automatically. This is simplifying the process of establishing automated work flows, but also the emerging data spaces.

3. Recommendations

Specific recommendations are listed below.

REQ-1 Always publish data at the highest possible functional granularity (i.e. not individual measurements, but neither several stations combined in one dataset).

REQ-2 Never combine data with different temporal dimensions (e.g. minute and hourly resolutions) in the same dataset.

REQ-3 Never combine data with different vertical dimensions (e.g. surface observations and vertical profiles) in the same dataset.

REQ-4 If there is a need to reference datasets to collection work (e.g. research cruises or field work), establish this reference through tags on the data or parent/child relations allowing data consumers to collect data based on content and spatio-temporal position.

[1] Details on how to avoid duplicate information in WIS and WIGOS needs to be defined.

[2] As of today, WMO GTS, but to be replaced by MQTT.