POLAR RESEARCH INFRASTRUCTURE NETWORK

Deliverable 4.4 Guidance on dataset granularity

V1.2, March 2025

www.eu-polarin.eu

Funded by the European Union

POLARIN: POLAR RESEARCH INFRASTRUCTURE NETWORK

Funding programme: Horizon Europe

Grant Agreement No.: 101130949

Project Start Date: 01/03/2024

Duration: 60 months

Coordinator: Alfred Wegener Institute, Germany

Document information	
Work Package	WP4 Improvement of data services and customised data products
Deliverable No	D4.4
Deliverable title	Guidance on dataset granularity
Version	V1.2
Dissemination level	☒ PU - Public ☐ PP - Restricted to programme partners ☐ RE - Restricted to a group specified by the consortium ☐ CO - Confidential, only for members of the consortium
Lead Beneficiary	CNR
Lead author	Giulio Verazzo
Contributors	
Contributing authors	Giulio Verazzo
Due date	31/02/2025
Delivery date	24.03.2025

Document history	
Creation Date	19/12/2024
Version	[V1.2]
Version Date	21/03/2025
Status	☐ Draft ☒WP lead approved ☒ Coordinator approved ☒ Executive Board approved
Status date	24.03.2025

TABLE OF CONTENTS

SUMMARY

Granularity, the level of detail in datasets, is crucial for machine processing and reuse in activities like satellite calibration and numerical modeling. Publishing data with finer granularity improves metadata discovery, simplifies dataset usage, and reduces unnecessary data points. Data producers can optimize datasets granularity with actions like tailoring dimensions to specific profiles or time series and avoiding combining datasets with different temporal or vertical resolutions, as it complicates processing and aggregation. Relationships between datasets can be maintained through tagging or parent/child structures, enabling seamless integration into larger networks. These practices streamline data workflows, enhance automation, and ensure usability over long time periods.

1. Introduction

The polar regions are vital components of Earth's system, crucial for regulating our climate and serving as indicators of climate change, human expansion, and resource exploitation. These regions are experiencing rapid ice loss and significant transformations in their oceans and land, with global repercussions that impact people in diverse ways. To ensure sustainable development in the Arctic and effectively protect the Antarctic, evidence-based policy recommendations are necessary. However, the remoteness and inaccessibility of the polar regions, combined with limited research infrastructure, present significant challenges. Research data is often fragmented and dispersed across various databases, lacking sufficient interoperability. To better understand and predict key processes in the polar regions, and to provide the evidence-based information needed to support the European Green Deal and EU Arctic policy, the polar research community requires access to world-class research infrastructure in these areas.

POLARIN is an international network of polar research infrastructures and services designed to address the scientific challenges of the polar regions. This network encompasses a diverse range of top-tier research infrastructures, including Arctic and Antarctic research stations, research vessels, icebreakers operating at both poles, observatories, data infrastructures, and repositories for ice and sediment cores. POLARIN aims to provide integrated, challenge-driven access to these infrastructures, facilitating interdisciplinary research on complex polar processes.

To this end, POLARIN will:

- Offer challenge-driven transnational access to a broad portfolio of research infrastructures.
- Enhance data accessibility by improving data availability and interoperability among data infrastructures.
- Provide virtual access to data and data services.
- Deliver data products for the scientific community and decision-makers.
- Train the next generation of polar researchers to effectively utilize these infrastructures for their research.
- Actively promote the services offered by POLARIN and encourage infrastructure users to share their research outcomes with society.

This deliverable aims to give guidance on dataset granularity to data producers in the context of POLARIN. It is based on the SIOS' *Granularity Perspectives Document* and CNR's best practices on datasets granularity.

2. Background

Granularity refers to the level of detail or depth in a dataset. This is a constant challenge in distributed data management. Frequently, the information available at the discovery level is suited for human interpretation but lacks the structure needed for machine processing. This limitation hinders the creation of aggregated datasets and prevents computers from handling data preparation, leaving humans less time to concentrate on interpretation and analysis.

Dataset granularity is often determined by the data provider (e.g., the data collector or generator), whose perspective on the data may differ from that of a potential data consumer. Providers may choose to publish data from a single project or research expedition as a unified collection, allowing for a single publication to cite. This approach is convenient for project participants to share data among themselves as a single collection, regardless of how it is ultimately published. However, when data is published, it should be viewed as part of a much larger data network.

Publishing data with finer granularity offers several benefits:

- Discovery metadata can be provided at a more detailed level, making it easier for users to locate and isolate the specific data they need.
- Each file may contain fewer dimensions, making it simpler to create, interpret, read, and build services around.
- NetCDF files are less likely to include excessive fill values, as their dimensions can be tailored to specific profiles or time series rather than accommodating all data at once.

One common concern is that downloading multiple files might complicate access for data consumers. One way to address this problem is the usage of an OPeNDAP-based repository to store data. This type of repositories leverages the OPeNDAP data serving protocol which allows to query specifying subsets of the dataset, filtering variables and their values. Moreover, it provides a “streaming-like” access to the data, removing the need to download a potentially large dataset.

Additionally, data centers could implement services to merge multiple files into a single file upon download. Although such services are not yet widespread, data should be published with consideration for future possibilities and the long-term timeframe over which the data is likely to be used—often spanning several decades.

3. Granularity in practice

In general, the highest possible granularity—where data is separated into the most detailed datasets—is advantageous for machine processing and reuse in applications such as numerical models, satellite calibration, and validation activities. This means avoiding the aggregation of multiple stations into a single dataset or file and instead keeping them as separate datasets. These individual datasets can be organized into networks, field activities, research cruises, etc., by establishing parent/child relationships or tagging them with specific keywords.

Another important principle is to avoid combining data recorded at different time scales within the same dataset. For instance, if a station has sensors observing data at both minute and hourly intervals, these should be published as two separate datasets. While combining them into a single dataset is technically possible, it increases complexity for downstream machine processing, adding unnecessary degrees of freedom for software tasked with data aggregation.

Similarly, feature types (e.g., time series and time series of profiles) should not be mixed within the same dataset. For example, weather stations often include instruments with distinct characteristics: some may observe scalar values at a specific height (e.g., air temperature), while others record profiles (e.g., permafrost). Publishing these as separate datasets minimizes complexity, making it easier for software to combine data efficiently.

The goal of these publishing restrictions is to streamline the process of integrating datasets into spatially and temporally aggregated collections automatically. This not only simplifies the development of automated workflows but also supports the creation of emerging data spaces.

4. Recommendations

Specific recommendations are listed below.

1. Always publish data at the highest possible functional granularity (i.e. not individual measurements, but neither several stations combined in one dataset).
2. Never combine data with different temporal dimensions (e.g. minute and hourly resolutions) in the same dataset.
3. Never combine data with different vertical dimensions (e.g. surface observations and vertical profiles) in the same dataset.
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.

Moreover, it's useful to take into consideration the following lists of best practices when creating datasets in CSV and netCDF formats:

For CSVs:

- Do not use the same name for different variables
- Use the dot as decimal separator
- It's ok to write the variable names in the first line and the unit measures in the second but it's not mandatory
- Use UTF-8 character encoding for variables names and unit measures
- Use decimal degree for the latitude in the range between: -90:90
- Use decimal degree for the longitude in the range between: -180:180
- Use the "time" name for the time variable, "latitude" name for the latitude, "longitude" for the longitude, "depth" for the depth and "altitude" for the altitude.
- Do not split date and time on two or more variables, use just one variable called "time" and write the timestamp as ISO 8601 format (yyyy-mm-ddThh:mm:ss) expressed as UTC

For netCDFs:

- Do not use the same name for different variables
- Use the dot as decimal separator
- Use UTF-8 character encoding for variables names and unit measures
- Use decimal degree for the latitude in the range between: -90:90
- Use decimal degree for the longitude in the range between: -180:180
- Use the data type *float* for latitude, longitude, depth and altitude and for floating point numbers
- Set *yyyy-mm-dd'T'hh:mm:ss* as the time unit.

5. Acknowledgements

POLARIN is a project that has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101130949. Please visit www.eu-polarin.eu for more information.