POLARIN

POLAR
RESEARCH
INFRASTRUCTURE
NETWORK

Deliverable 5.1. A unified semantically consistent virtual data catalogue with machine interfaces

V1, 28-02-2025

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	WP 5 Provision of virtual access
Deliverable No	D5.1
Deliverable title	A unified semantically consistent virtual data catalogue with machine interfaces
Version	[Final]
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	SIOS
Lead author	Daniël Kivits
Contributors	SIOS
Contributing authors	Ilkka Matero, Øystein Godøy, Lara Ferrighi
Due date	2025-03-01
Delivery date	2025-02-28

Document history	
Creation Date	2025-02-08
Version	[V1]
Version Date	2025-02-28
Status	☐ Draft ☒ WP lead approved ☒ Coordinator approved ☒ Executive Board approved
Status date	2025-02-28

TABLE OF CONTENTS

EXECUTIVE SUMMARY

The development of a unified, semantically consistent virtual data catalogue is a key step in facilitating seamless access to research data collected at research infrastructures offering Transnational Access and Virtual Access in POLARIN. This deliverable outlines the progress made in establishing a federated data catalogue that integrates existing discovery metadata standards and ensures interoperability between regional Arctic and Antarctic data infrastructures. The deliverable outlines the technical foundation, current implementation progress, and future steps toward full integration into the POLARIN web portal. By leveraging existing brokering solutions and regional interoperability frameworks, this effort contributes to a more cohesive European and international polar data landscape. By leveraging existing brokering solutions and internationally adopted interoperability frameworks, this effort contributes to a more cohesive European and international polar data landscape.

1. Introduction

1.1. Background

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 significant changes in ice cover, 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. Furthermore, the remoteness and inaccessibility of the polar regions, combined with limited research infrastructure (RI), present significant challenges. Research data is often fragmented and dispersed across various databases thatlack 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 will benefit from access to world-class RI in these areas.

POLARIN is an international network of polar RI and services designed to address the scientific challenges of the polar regions. This network encompasses a diverse range of top-tier RIs, 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:

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

The accessibility and integration of polar research data remain crucial challenges in advancing scientific collaboration and decision-making. RIs across the Arctic and Antarctic generate vast amounts of data, yet these datasets are often published within independent repositories, each with distinct metadata standards and exchange protocols. To address the projects' data-related objectives (second to fourth in the list above), POLARIN aims to establish a unified virtual data catalogue, providing a single-entry point for researchers, policymakers, and other stakeholders to discover and access datasets from participating infrastructures. The POLARIN data catalogue will provide access to diverse data sources in a semantically consistent and interoperable manner.

1.2. Definitions

1.2.1. Dataset

POLARINs Virtual Access will contain a variety of *data products*, ranging from raw data to processed and aggregated data. Among others, it will contain point data, trajectory data, profile data, time series, and gridded data. The term *data product* is quite ambiguous, and therefore the term *dataset* will be used to refer to it in the rest of the document. This can be observational data but also processed observations for a specific purpose. The working definition of a dataset in this context is in line with the INSPIRE directive (Mäkelä, 2007):

"… an identifiable collection of spatial data, i.e. a collection of data that has a reference by name or coordinates to a geographic location or area, and which in addition have a designated start and end time."

A dataset can contain observations (remote or in-situ), derived quantities (from either of these two types of data sources), or forecasts of future states of environmental parameters. Data values can be located at a single point, along a line or transect, in a regular or irregular grid, and be captured or estimated at one or more altitudes/depths. A dataset can be stored on paper, in files (one or more), or in a database, and is often accompanied by descriptions (metadata) of its content.

1.2.2. Discovery metadata

Discovery metadata describes the 'who', 'where', and 'when' of the data collection process, how to access data and any potential constraints on the data. It should also link to further information on the data, if relevant. It includes information like titles, keywords, abstracts, geographic locations, dates, and creators. Its primary role is to make resources discoverable in catalogs, repositories, or search engines.

1.2.3. Metadata standard

A metadata standard is a set of rules and guidelines that define how metadata should be created, formatted, and exchanged between different systems. Metadata standards provide a common language and structure for metadata, ensuring that different systems can understand and interpret the metadata correctly. By following a metadata standard, developers can ensure that their metadata is consistent and easily consumable by other systems. Examples of discovery metadata standards are ISO-19115, GCMD DIF and Attribute Convention for Data Discovery (ACDD).

1.2.4. Metadata exchange protocol

The use of standardized metadata enables the use of metadata exchange protocols, often referred to as 'machine-to-machine (M2M) interfaces', which are vital for automating data discovery, integration, and analysis. The use of such metadata exchange protocols is key to support the harvesting procedures of POLARIN, enhancing usability and enabling complex, automated workflows across various scientific domains. Examples of metadata exchange protocols are OGC CSW, OAI-PMH, and OpenSearch.

1.3. Scope

The POLARIN data catalogue has been built on experience gained from building other existing discovery metadata brokering approaches available within the consortium (SIOS, INTERACT), other European projects (Arctic PASSION, ENVRI-FAIR, EOSC) and within the broader polar data

management community (POLDER, WMO Global Cryosphere Watch). The state-of-the-art POLARIN data catalogue will be part of the backend of the POLARIN web portal (see Deliverable 4.3), which will focus on providing both human-friendly and machine-readable access, and enabling seamless data integration into analytical workflows and decision-support systems. The POLARIN data catalogue focuses on discovery information from POLARIN RIs and the ability to enrich and group these records.

Because the POLARIN data catalogue and POLARIN web portal are closely linked, this document not only outlines the main objectives, the technical approach and the implementation progress of the POLARIN data catalogue itself, but also provide an outlook on the future developments necessary to achieve a fully operational and user-centric POLARIN web portal.

2. Main objectives

The objectives of this deliverable, as stated in the POLARIN Description of the Action, are:

- Establish a unified virtual data catalogue giving access to data from all POLARIN RIs offering VA.
- Providing backend functionality for Work Package (WP) 4
- Enabling brokering of regional European and Polar data management frameworks at discovery level

3. Technical approach

The POLARIN data catalogue is refered to in the POLARIN Description of the Action as ‘a unified semantically consistent virtual data catalogue with machine interfaces’. This chapter will describe the architecture of the POLARIN data catalogue in more detail, and describe how the data catalogue was created.

3.1. Architecture overview

The POLARIN data catalogue is served through machine endpoints that are implemented using pyCSW. In order to feed the catalogue harvesting, transformation and filtering of harvested discovery metadata records is supported using in-house developed software in Python (GPL Licensed). In the transformation of information to the internal data model and in the filtering steps, information can be enriched. In this process no actual data are moved, just information about the data. Data are served from the host data repository.

An overview of the architecture of the POLARIN data catalogue is given in Figure 1. Here is a short overview of the different parts of the data catalogue setup shown in that figure:

Table 1: Components of the POLARIN data catalogue architecture

(md)harvester	this is where all discovery metadata records from the data repositories ('external sources' in the schematic) are harvested. It connects to a number of standard Application Programmers Interfaces and extracts a local copy of the metadata records available.
mdtransform	this is where the metadata documents provided by data repositories (i.e. the metadata harvested in the previous step) are translated to the MET Norway Metadata Format (MMD) supported by a vocabulary service (https://vocab.met.no/). The latter is linked to other vocabulary services, e.g. the NERC Vocabulary Server, the Global Change Master Directory keywords vocabulary, as well linked to relevant international resources, e.g. the WMO Space-based Capabilities (OSCAR/Space) catalogue. In the translation process of the metadata harmonisation and enrichment of the information is possible. This is done through semantically constistent mapping of harvested metadata.
mdfilter	this is where POLARIN relevant metadata are filtered on e.g. variable content, geographical location, are checked on uniqueness, and where harvested metadata records can be connected to RIs if not done by the provider.
mdingest	this is where the filtered metadata records are ingested into SolR (search engine). SolR is the backend of the mdcat (machine readable interface – pyCSW).
mdcat	this is the machine-readable interface of the POLARIN data catalogue, that offers M2M communication through a pyCSW implementation. This includes (but is not limited to) OAI-PMH, OGC CSW, and OpenSearch.

Any discovery metadata that are exposed via the POLARIN data catalogue follows the same information flow: metadata records are harvested from multiple external resources; the metadata records are then translated into an internal metadata model; the translated metadata records relevant to POLARIN are identified and tagged; the filtered metadata records are then ingested into a central database / search engine and then exposed in a standard machine-readable interface.

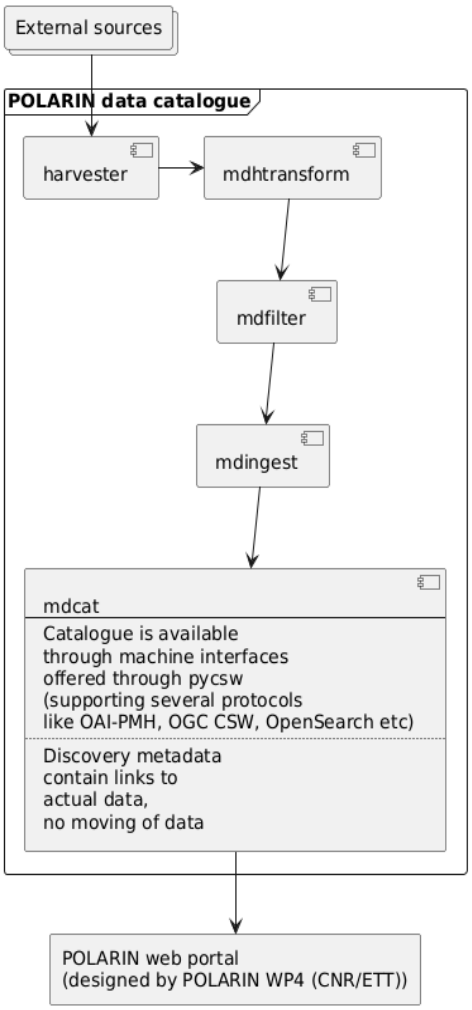

Figure 1: POLARIN data catalogue architecture diagram

3.2. Connection to POLARIN web portal

For deliverable *D5.2: "A web portal providing guidance documentation and a graphical user interface to the virtual data catalogue"* the POLARIN data catalogue will be connected to the POLARIN web portal designed under deliverable *D4.3: "POLARIN web portal".* This means that the standalone POLARIN data catalogue (described in this report) and the POLARIN web portal (described in D4.3) solutions will be combined into one solution for D5.2, to create an integrated POLARIN web portal: a virtual access portal that provides both human-friendly and machine-readable interfaces to POLARIN reasearch data, with the goal to enable researchers to discover, access, and utilize POLARIN-related datasets efficiently.

4. Implementation status

The POLARIN data catalogue is ready to be populated with research datasets relevant to POLARIN and the machine-readable interface is accessible via the following link: http://polarin-sios.csw.met.no/.

The POLARIN data catalogue is hosted by the Norwegian Meteorological Institute (MET), who provide the data management services for SIOS. They host numerous research datasets relevant to POLARIN through the SIOS Data Management System (SDMS), that also feed into other regional European and Polar data management frameworks (mostly at the discovery level). All of the metadata records of these datasets are managed in a semantically consistent manner, as described in more detail in Section 3.1.

The POLARIN data catalogue offers the OGC CSW and OAI-PMH metadata protocols for machine-interoperable metadata access, and while OpenSearch is also being supported, there is more development needed to ensure full support of all queries. However, pyCSW allows for efficient implementation of emerging machine-to-machine (M2M) protocols, making it future-proof for evolving standards and specialized needs. Overall, pyCSW's support for multi-protocol, machine-readable interfaces strengthens data accessibility and fosters collaboration, ensuring continued relevance as data-sharing standards evolve. Within 2025 a dedicated development effort with members of OSGeo is planned to improve the abstraction level of pyCSW. This will also add support for OGC API Records and Spatio-temporal Assets Catalogue (STAC) when running pyCSW on top of SolR or ElasticSearch.

5. Future work

To populate the POLARIN data catalogue with relevant metadata records, we first need to identify the research datasets relevant to POLARIN. To achieve this, we first need to gain understanding of how the various RIs behave with regards to their research data, where their research data is currently published, and of the quality of the published records.

Then, we need to work with the RIs and the data repositories they are affiliated with to improve tagging of datasets. Such tagging should reflect both the link to the research infrastructure in question and POLARIN. Through WP6, POLARIN can offer (limited) support to POLARIN RIs that need extra resources to improve the tagging of their datasets.

6. Summary

We have developed a unified POLARIN data catalogue that is ready to expose discovery metadata describing research data relevant to POLARIN. The POLARIN data catalogue supports several discovery metadata standards and exchange protocols. The POLARIN data catalogue is accessible via the following link: http://polarin-sios.csw.met.no/.

The catalogue will be populated after we have analysed past, current and future behaviour of RIs in the field of data management. We will guide POLARIN RIs on how to connect with a data repository if needed, and provide (limited) support to POLARIN RIs to upload their data to these repositories if needed.

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.

References

POLARIN Description of the Action (Associated with document Ref. Ares(2023)7293125 – 26/10/2023)

Mäkelä, T. (2007). *INSPIRE Directive of the European Parliament and the Council establishing an Infrastructure for Spatial Information in the European Community*. Official Journal of the European Union, L 108/1.