ARCTIC PASSION

Deliverable 2.1

***Data Management Plan***

Version 0.4, 2023-04-19: First update

**Project**
Arctic PASSION

**EU Horizon 2020 grant agreement**
101003472

**Work package 2**
Bringing the Arctic Data System into action

**Lead beneficiary**
1 - FMI

**Lead author**
Matias Takala (FMI)

**Contributors**
Øystein Godøy (MET)

**Status**
Coordinator accepted

**Dissemination level**
PU

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101003472

# Table of Contents

# 1. Data summary

The purpose of the data management plan is to document how the data generated by the project Arctic PASSION are handled during and after the project. It describes the basic principles for data management within the project. This includes standards for documentation at discovery and data levels as well as data sharing and preservation including life cycle management of datasets.

| | |
|---|---|
| **IMPORTANT** | This document is addressing datasets (e.g. observations of conditions in the ocean, atmosphere, cryosphere). In addition the project is collecting information on activities that are generating data like research cruises, field work activities etc. This type of information is not covered by this document, but the information model used for the data catalogue allows linking data with further information about the mechanisms used to collect data. |

This document is a living document that will be updated on a regular basis during the project. Updates will be made when deemed necessary, but updates will at least be made in relation with the Arctic PASSION reporting periods, which are currently scheduled for December 2022, June 2024 and June 2025. Arctic PASSION is following the principles outlined by the Open Research Data Pilot and The FAIR Guiding Principles for scientific data management and stewardship (Wilkinson et al. 2016). The data management plan is based on the OpenAIRE guidelines [https://www.openaire.eu/how-to-create-a-data-management-plan].

The Arctic PASSION project is described in more detail in the project website which is available at https://arcticpassion.eu/. The purpose of Arctic PASSION is creation and implementation of a coherent, integrated Arctic observing system. This implies integrating already existing observing systems in a systems of systems approach as well addressing gaps in the current observing system. Arctic PASSION will generate data through the project implementation, but equally important is to map and establish access to already existing datasets in support of the pilot services and other activities of Arctic PASSION.

Arctic PASSION will promote the use of self-explaining file formats (e.g. NetCDF, HDF/HDF5, DwCA) combined with semantic and structural standards like the Climate and Forecast Convention for data documentation. The default format for Arctic PASSION datasets in the geoscientific domain is NetCDF following the Climate and Forecast Convention (feature types grid, timeseries, profiles and trajectories if applicable). For data in the biological domain Darwin Core Archive is promoted. If none of these formats are suitable other formats can be used, but a detailed product manual following a template has to be prepared to ensure proper reuse of the data in the future.

Arctic PASSION will exploit existing data in the region. In particular operational meteorological data made available through WMO Global Telecommunication System (GTS) will be important for the model experiments. No full overview of third party data that will be used is currently available. An overview of the third party data that are planned to be used by the pilot services and that need some sort of handling within the Arctic PASSION data catalogue will be provided in subsequent updates of this plan based on input from the pilot services. Essentially this will e.g. include data from the World Meteorological Organisation, Copernicus services in Europe, data already generated by project partners and data found when harvesting discovery metadata from relevant data centres.

If deemed necessary (required by the scientific community in Arctic PASSION) metadata describing relevant third-party observations will be harvested and ingested in the data management system and through this simplifying the data discovery process for Arctic PASSION scientists. If specifically needed by one of the pilot services of Arctic PASSION, data may also be cached to ensure interoperable data that can be used by the web based services of the pilot services[1].

Arctic PASSION will rely on data generated by project partners during the duration of the project, legacy data and observing systems of the partners and third party data available through data centres not part of Arctic PASSION.

An overview of the data generated (or used) by the project will be made available in Appendix A, more specifically in Table 1 which serves as a reminder for datasets the project is generating and Table 2 which lists datasets already available, but that Arctic PASSION is planning to actively use in services. Information in Table 1 is primarily building on existing measurement programmes or extending these to new areas. Eventually all datasets should be discoverable through the Arctic PASSION data catalogue.

| | |
|---|---|
| **IMPORTANT** | Table 1 is vaguely populated and will be populated in more detail in the subsequent versions of the data management plan following interaction with the data generating work packages (in particular WP 1 and 3) of Arctic PASSION. There is currently no estimate for the expected volume of the data. Such volume estimates only make sense for the data actively managed by Arctic PASSION. These estimates will be generated when a better overview of the exact datasets is available. However it is expected that it will be in the order to several Terabytes. |

Arctic PASSION aims to *bring the Arctic data into action*. Thus data can be relevant for many communities. Internally the primary purpose of the data is to serve the needs of the project's pilot services.

# 2. FAIR data

## 2.1. Making data findable, including provisions for metadata

Arctic PASSION will use the SAON data portal (Figure 1), for the time being accessible through https://saon.met.no/ [2], to serve data consumers with both human and machine interfaces. Human and machine interfaces relies on a data catalogue that is generated using an information mode that is in use for multiple projects and activities. This is the MET Norway Metadata Format Specification (MMD) [https://htmlpreview.github.io/?https://github.com/metno/mmd/blob/master/doc/mmd-specification.html]. This is developed to be compliant with GCMD DIF and ISO19115 and is widely used for mapping harvested metadata into a unified data model. Mappings to DCAT is in progress.

The SAON Data Portal doesn't host datasets, but harvest information about datasets from a number of data repositories[3] and integrates this information in a unified search interface.

| | |
|---|---|
| **IMPORTANT** | Arctic PASSION require that data generated in the project are published in a data |

centre that allows machine access using standard interfaces and information objects for discovery metadata.

**IMPORTANT**

Arctic PASSION require that data generated in the project are published with a project tag populated with the text "Arctic PASSION" in both the short and long name for the project. This is used to group Arctic PASSION datasets in the SAON Data Portal.

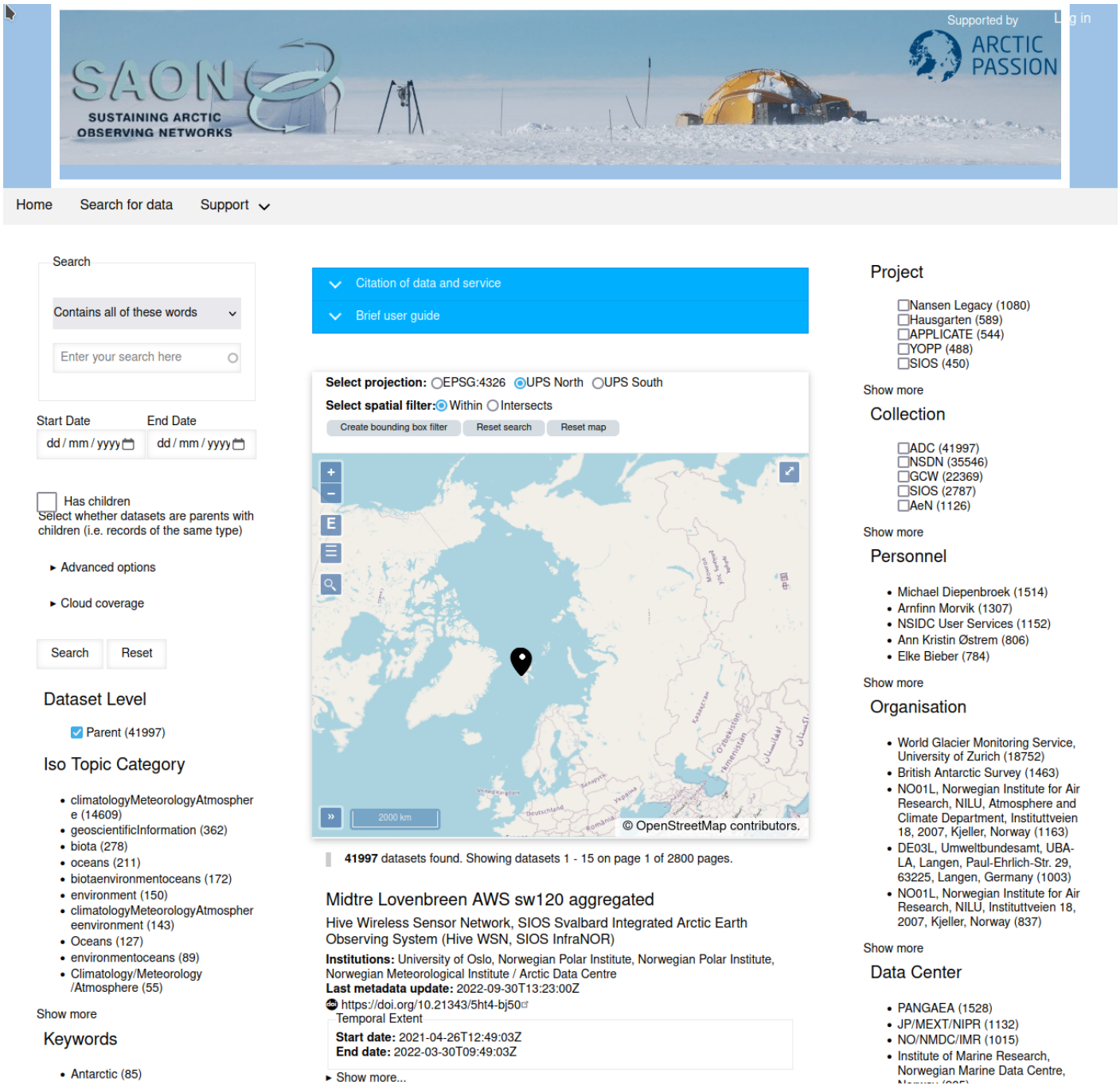


*Figure 1. Screenshot of the SAON Data Portal (for the time being available through https://saon.met.no/) which Arctic PASSION supports the development of.*

**NOTE**

Although Arctic Council has suspended all *official meetings of the Council and its subsidiary bodies until further notice*, no information is received to suspend operation of

> the SAON Data Portal. The SAON Data Portal is an in kind contribution from the Norwegian Meteorological Institute.

When data are served using self-describing file formats like NetCDF according to the Climate and Forecast Conventions [https://cfconventions.org] with global attributes according to the Attribute Convention for Dataset Discovery [https://wiki.esipfed.org/Attribute_Convention_for_Data_Discovery_1-3] (ACDD)[4] and served through OPeNDAP, discovery metadata can be directly generated from the data files. A similar set up is possible to achieve with Darwin Core Archives [http://tools.gbif.org/dwca-assistant/] (DwC-A), which also have metadata embedded. However, the procedure for extracting this information is yet not operational in the context of Arctic PASSION. The workflow for CF-NetCDF is currently in testing. The workflow for DwC-A is still under development. In essence application of CF-NetCDF and DwC-A addresses both the perspectives of making data findable and interoperable.

| | |
|---|---|
| **IMPORTANT** | Sensitive data generated by community based monitoring will be handled in a separate system and only aggregated information will be made available in the data catalogue. However, this data Management Plan will also be developed to cover the sensitive data. |

## 2.2. Making data openly accessible

Data will be served from the host data centre wherever possible. Datasets that are needed by a pilot service, but are not openly available although the data license allows open access, will be cached by MET during the project duration and made available for potential users internally and externally.

Selected datasets are preserved for the future through PANGAEA and FMI who will also provide discovery metadata and online access to these datasets.

MET offers limited (large volumes may be too costly) hosting support for "homeless data" that are important for the project deliverables. If data providers have funding to support hosting of large datasets, this can be discussed with MET.

## 2.3. Making data interoperable

Arctic PASSION will primarily rely on self describing, standardised file formats for data encoding. These standardised formats also have semantic frameworks for annotation of the data. This simplifies integration of data across data providers and communities and is in line with efforts undertaken in large data exchange activities, like operational data exchange through the World Meteorological Organisation (WMO) working with atmospheric, oceanographic and hydrological data and the Global Biodiversity Information Facility [https://www.gbif.org/] (GBIF). The specific standards that will be promoted by Arctic PASSION include:

**CF-NetCDF**

NetCDF adhering to the Climate and Forecast Conventions [http://cfconventions.org/index.html] is widely used, both in the oceanographic community, in the Earth System Grid Federation, in Copernicus services, by ESA and EUMETSAT for Sentinel data provision and WMO is developing WMO specific

profiles of the standard. By adding the Attribute Convention for Dataset Discovery [https://adc.met.no/node/4][4], discovery level metadata can be embedded in the datasets.

**Darwin Core Archive**

According to the Darwin Core Archive Assistant [http://tools.gbif.org/dwca-assistant/] *Darwin Core Archive (DwC-A) is a Biodiversity informatics data standard that makes use of the Darwin Core terms to produce a single, self contained dataset for species occurrence or taxonomic (species) data. It is the preferred format for publishing data to the Global Biodiversity Information Facility. You export your data as a set of one or more text (CSV) files. A simple XML descriptor file (called meta.xml) is required to inform others how your files are organized.*

Data that doesn't fit into these categories will be accompanied by a detailed product manual providing guidance to data consumers. These data will require some more human effort to utilise. Both CF and DwC-A standards are managed in well defined governance processes and the standards are used widely beyond the original user communities.

| | |
|---|---|
| **IMPORTANT** | The template for the product manual is to be developed. |
| **IMPORTANT** | Guidance on how to use the standards mentioned above will be made available through https://saon.met.no/apguidance. |

# 2.4. Making data reuseable

A very important requirement for reuseable data is that data are released using a clear data license. Arctic PASSION will promote the usage of the Creative Commons Attribution 4.0 International [https://spdx.org/licenses/CC-BY-4.0.html] license.

The standards for use metadata that are promoted by Arctic PASSION, i.e. Climate and Forecast Conventions [http://cfconventions.org/index.html] and Darwin Core [https://www.gbif.org/darwin-core] ensures self describing data according to a shared terminology.

As noted in the previous chapter, not all data fits in these formats. These data will not follow rich metadata standards and will require human effort to properly reuse.

When data are documented according to the standards mentioned above, reuse is simplified as standardised tools and services will offer support out of the box. CF-NetCDF and DwC-A is e.g. widely used within many data exchange frameworks.

While CF-NetCDF have been widely used in many communities for a long time, the standard is pretty wide and the degrees of freedom sometimes makes it hard to maintain software support for all options, not least when integrating data across providers. WMO has recognised this and trough interaction with the CF governance, WMO has included CF-NetCDF as part of the WMO Information System [https://public.wmo.int/en/wmo-information-system-wis] (WIS) governance through a dedicated Task Team on CF-NetCDF [https://community.wmo.int/governance/commission-membership/commission-observation-infrastructure-and-information-systems-infcom/commission-infrastructure-officers/infcom-management-group/standing-committee-information-management-and-technology-sc-imt/expert-team-data-standards-1] which will develop WMO profiles of

the CF standard for specific WMO purposes.

# 3. Allocation of resources

Arctic PASSION Work Package 2, Bringing the Arctic Data System to action, has allocated resources for cataloguing, serving and preserving data within the project period. Handling of sensitive data from Community Based Monitoring is done in Work Package 4. Overall responsibility for the Data Management Plan lies with Work Package 2.

# 4. Data security

Most of the data generated by Arctic PASSION is open. Arctic PASSION is working to establish secure connections between data centres and data consumers to ensure that correct decisions can be made using data. However, data from third parties will also be made available, for these data there is limited room for Arctic PASSION to ensure integrity and security of data.

| | |
|---|---|
| **IMPORTANT** | Arctic PASSION promotes the application of secure transport protocols between data centres and data consumers. |
| **IMPORTANT** | For the discovery metadata harvested into the Arctic PASSION data catalogue, translation rules have been developed that rely on well defined document standards and controlled vocabularies/terminologies. This is further described in the project deliverable (D2.3) which describes the website. |

Data from Community Based Monitoring that could be of sensitive nature will not be publicly available, only aggregated non sensitive information will be available through the Arctic PASSION data catalogue.

# 5. Ethical aspects

As mentioned above, sensitive information from Community Based Monitoring is handled in a separate system adhering to the ethical and legal regulations for such data. There could be other information that has constraints for ethical reasons (e.g. species information or breeding areas), but identification of these will be part of the further development of the data management plan and in particular Table 1.

| | |
|---|---|
| **IMPORTANT** | Data within Arctic PASSION will be handled according to the principle of "as open as possible, as closed as necessary". |

# 6. Other issues

A major challenge when working with scientific communities is to raise the awareness of interoperability at the data level. Often data are published and shared in the form of spreadsheets or in other unstructured forms, which complicates efficient reuse of the data in decision support systems. Arctic PASSION is actively working to change this, but it is a task that is tedious and time consuming

since the cost for scientists to overcome the threshold of using FAIR compliant file formats is substantial and the benefit is not evident immediately.

# 7. References

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

# Appendix A: Datasets

*Table 1. Overview of datasets generated within Arctic PASSION. Dataset definitions are preliminary and high level. Each record will materialise in many discovery metadata records.*

| # | Dataset | Description | Responsible | Generated | Published | Comment |
|---|---|---|---|---|---|---|
| 1 | CTD-data | CTD casts taken during regular cruises in the Arctic and surrounding areas | | | | Details are still being investigated. |
| 2 | Mooring-data | Information from long-term ocean moorings of temperature, current etc. | | | | Details are still being investigated. |
| 3 | CBM climate data | Climate information from Community Based monitoring | | | | Details are still being investigated. |
| 4 | Aerosol-data | Information on Arctic Aerosols. | CNR | | | Details are still being investigated.. |
| 5 | Surface irradiance measurements | Information on the short- and longwave surface irradiance. | | | | Details are still being investigated. |
| 6 | Surface weather stations | | | | | Details are still under investigation, based on relations to INTERACT |

| # | Dataset | Description | Responsible | Generated | Published | Comment |
|---|---|---|---|---|---|---|
| 7 | Terrestrial data | Information on terrestrial features, including biodiversity and snow etc. | | | | Details are still being investigated. |
| 8 | Permafrost data | Depth profiles of temperature in the permafrost. | | | | Details are still being investigated. |
| 9 | Ice mass balance buoys | | | | | Details are still being investigated. |
| 10 | In situ observations of sea ice | Information received from ships in the ice through the IceWatch activity. | | | | Details are still being investigated. |
| 11 | Arctic Land Ice from satellite | | | | | Details are still being investigated. |
| 12 | Airborne snow and ice data | | | | | Details are still being investigated. |
| 13 | Ice-Thethered Ice Observatories | | | | | Details are still being investigated. |
| 14 | Numerical simulations | Supporting observation impact studies, including climate at different temporal scales. | | | | Details are being investigated. |
| 15 | | | | | | |

| # | Dataset | Description | Responsible | Generated | Published | Comment |
|---|---|---|---|---|---|---|
| 16 | | | | | | |

*Table 2. Overview of datasets to be actively used by Arctic PASSION services.*

| # | Dataset | Description | Responsible | Generated | Published | Comment |
|---|---|---|---|---|---|---|
| 1 | TOPAZ (ARC MFC) | Sea ice concentration and thickness forecast | MET Norway, NERSC | Daily | CMEMS | Daily 10 day forecast |
| 2 | NeXtSIM (ARC MFC) | Sea ice concentration and thickness forecast | NERSC | Daily, Monthly | CMEMS | 1-day hindcast and 9-day forecast |
| 3 | GLO MFC | Sea ice concentration and thickness forecast | Mercator Océan International | Daily, Monthly | CMEMS | |
| 4 | Baltic MFC | Sea ice concentration and thickness forecast | SMHI | Sub-hourly, Hourly, Daily, Monthly | CMEMS | 6-day forecast |
| 5 | EUMETSAT OSI SAF time series | Sea ice extent/area | MET Norway | Daily | EUMETSAT | |
| 6 | Baltic OMI (CMEMS) | Sea ice extent/area | Ifremer (FMI/SMHI) | Daily | CMEMS | |
| 7 | Surface weather stations | Information from operational surface weather stations where data is exchanged through programmes of WMO | MET | | | Extraction of public available information is in progress |
| 8 | Vertical profiles of temperature | Information from operational stations exchanging data through programmes of WMO | MET | | | Extraction of public available information is in progress |
| 9 | | | | | | |
| 10 | | | | | | |

| # | Dataset | Description | Responsible | Generated | Published | Comment |
|---|---|---|---|---|---|---|
| 11 | | | | | | |
| 12 | | | | | | |
| 13 | | | | | | |

[1] This could be necessary to establish an Arctic Window of Copernicus or when data are available through third party data centres but not in standardised and interoperable form.

[2] This address will change at some point to the internet domain of SAON

[3] Details to be provided.

[4] More detailed information on how to format the ACDD global attributes to ensure the best possible discovery metadata being generated is available at https://adc.met.no/node/4.