ARCTIC PASSION

Deliverable 2.1

Data Management Plan

Version 0.2, 2022-03-24: Draft version

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
In preparation

Dissemination level
PU

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.

This document is a living document that will be updated during the project. 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 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 to be used will be provided. Essentially this will at least, but limited to, include data from the World Meteorological Organisation, Copernicus services in Europe, data already generated by project partners and data found 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 both 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 by the project will be made available in Appendix A, more specifically in Table 1. This list serves as a reminder for datasets. Eventually all datasets should be discoverable through the Arctic PASSION data catalogue.

IMPORTANT	Table 1 will be populated in the next version of the data management plan.

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, accessible through https://saon.met.no/, 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.

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)[2] 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 still not operational in the context of Arctic PASSION. 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 is 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 includes:

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][2], 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 use metadata standards 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 relies on well defined document standards and controlled vocabularies/terminologies. This is further described in the project deliverable on the website.

Data from Community Based Monitoring that could be of sensitive nature will not be public 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 from 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

None known yet.

Appendix A: Datasets

Table 1. Overview of datasets generated by Arctic PASSION.

#	Dataset	Description	Responsible	Generated	Published	Comment
1						
2						
3						
4						
5						
6						
7						
8						
9						
10						
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] 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.