First version of Data management plan for data from TNA activities

Work package n°	5
Deliverable n°	5.2
Lead beneficiary	ICOS ERIC
Author(s)	Damien Boulanger, Guillaume Brissebrat, Richard Rud, Markus Fiebig, Lynn Hazan, Cathrine Lund Myhre, Ute Karstens, Claudio D'Onofrio, Valérie Thouret, Pawel Wolff, Alex Vermeulen
Deliverable Type	ORDP: Open Research Data Pilot
Dissemination Level	Public
Estimated delivery date	M6
Actual delivery date	14 February 2022
Version	2
Reviewed by	WP5 Task leaders
Accepted by	ATMO-ACCESS Project office
Comments	*This initial DMP was created using DMPonline (https://dmponline.dcc.ac.uk/). It is a living document and will be updated by M30 of the project.*

This work has received funding from the European Union's Horizon 2020 research and innovation programme through the ATMO-ACCESS Integrating Activity under grant agreement No 101008004

Contents

1. ATMO ACCESS Project

A Data Management Plan created using DMPonline (https://dmponline.dcc.ac.uk/)

Creator: Alex Vermeulen

Affiliation: ICOS ERIC

Funder: European Commission

Template: Horizon 2020 DMP

ORCID iD: 0000-0002-8158-8787

Project abstract:

The ambition of ATMO-ACCESS is to address the needs for developing sustainable solutions based on the principles of open access and to develop guidelines and recommendations for governance, management and funding for efficient and effective access provision suited to distributed atmospheric RIs. This project investigates the most suitable mechanisms that could lead to the sustainable provision of access to atmospheric research infrastructures.

The main objectives of ATMO-ACCESS are:

- to provide coordinated open physical, remote and virtual access to state-of-the-art facilities and services in atmospheric RIs and further enhance their range of products, capabilities and accessibility for a wide range of users, including the private sector

- to engage facilities and their national stakeholders and direct them towards improved harmonisation of access procedures across the different member states, while also exploring modalities by which the use of atmospheric RIs can be further enhanced

- to explore and test new modalities of access that build on the complementarity and synergies among atmospheric RIs and respond to the evolving needs of users in relation to training, research and technology development, innovation and data services

- to identify the most suitable conditions for establishing sustainable access procedures across the EU for distributed atmospheric RIs, involving national and international stakeholders.

ATMO ACCESS
Access to Atmospheric Research Facilities

This project has received funding from the European Union's Horizon 2020 research and innovation programme through the ATMO-ACCESS Integrating Activity under grant agreement No 101008004.

ID: 88595

Start date: 01-04-2021

End date: 31-03-2025

Last modified: 14-02-2022

Grant number / URL: 101008004

2. ATMO ACCESS - Initial DMP

2.1.1. 1. Data summary

Provide a summary of the data addressing the following issues:

- **State the purpose of the data collection/generation**
- **Explain the relation to the objectives of the project**
- **Specify the types and formats of data generated/collected**
- **Specify if existing data is being re-used (if any)**
- **Specify the origin of the data**
- **State the expected size of the data (if known)**
- **Outline the data utility: to whom will it be useful**

In ATMO ACCESS different Research Infrastructures are involved that each represent different communities in the area of atmospheric research. Each of these have activities that generate data according to specific community standards that vary from RI to RI. Work from previous projects like ENVRIplus and ENVRI-FAIR have resulted in significant convergence of the FAIRness of the data handling and data repositories across the RIs. However, because of the different technical requirements and harmonization at the global level for their respective field of research each community still needs to support a set of unique data vocabularies, data types and data formats.

In the ATMO ACCESS project a lot of data will be generated that is specific to one or more of the RIs, next to more generic and/or common data. RI specific data of course will be curated by the most appropriate RI and stored at the respective data store of that RI. For the data management of this RI specific data the DMP or data lifecycle description of the responsible RI will apply, which should not be duplicated here.

The users of the Access services of ATMO ACCESS will generate data as well. The project will offer to host that data through the most applicable RI data center or data centers, in which case the respective RI DMPs will apply again. An important condition for all data provided through the ATMO ACCESS project is that the attribution to the project including the Project name and contract number is assured in the metadata.

The focus of this DMP is the generic and common data generated by the project. Curation of the data generated by the (virtual) access setup and managed by the project is part of the homeless data service developed in the project. All data curation there will be following the DMP of the respective RIs that will host the data when this data fits their capabilities. All other data not fitting the supported data categories will be referred to curation through more general repositories like Zenodo and Pangaea.

Applications for access to research facilities offered in the TNA and VA sections of the ATMO ACCESS project are open to the scientific community and will be evaluated and selected on the basis of scientific excellence. Due to this open character the type, size and timing of the (meta)data generated by the projects that will be supported is completely unknown and thus hard to describe in a single initial DMP and therefore this DMP will be extended in time through two additional successions over the course of the project as it becomes more clear which types and sizes of data will be generated.

It is however clear that we will support the community to deliver their (meta)data according to the FAIR principles as much as possible and through the services offered in the project for the homeless data most of the data to be expected will flow through the RI repositories and follow the FAIR workflow developed there.

The Virtual Access services will be developed in the first half of the project and will become operational in the last two years of the project. Of these two services will deliver mainly time averaged timeseries of atmospheric concentrations of atmospheric composition. These timeseries data sizes are usually in the order of single to maximally hundreds of megabytes for 10-20 year time series, so storage size will not be a big issue for either VA services or the TNA campaigns with their relatively short length in time. In the case that raw data needs to be stored on for example ceilometer or PTRS data, data volumes can increase to the order of several Gigabytes per campaign. Overall we can expect that the total data volume that will be generated in the ATMO ACCESS project will be maximally several TB, divided over the 3 repositories.

In principal all data generated by the project will be open and be accessible through open licenses such as CC BY 4, and shared through open repositories such as the repositories from the RI involved in the project, Zenodo, Pangaea/ Data from the TNA campaigns or data generated in the VA services can be part of other research projects and the data generators/owners will also have the need to publish the data under different licenses or keep part of the data under embargo until the research has been published. All RIs and repositories advised will support the possibility of embargo for the data and whenever possible offer as much as possible support for other licenses than the default of the repository.

Software generated specifically for the project will be shared as open source and made available through the GPL v3 licence model in an open repository like Github.

2.1.2. 2. FAIR data

2.1 Making data findable, including provisions for metadata:

- **Outline the discoverability of data (metadata provision)**
- **Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?**
- **Outline naming conventions used**
- **Outline the approach towards search keyword**
- **Outline the approach for clear versioning**
- **Specify standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how**

Applications for access to research facilities offered in the TNA and VA sections of the ATMO ACCESS project are open to the scientific community and will be evaluated and selected on the basis of scientific excellence. Due to this open character the type, size and timing of the (meta)data generated by the projects that will be supported is completely unknown at this time and thus hard to describe in a single initial DMP and therefore this DMP will be extended in time through two additional successions over the course of the project as it becomes more clear which types and sizes of data will be generated.

It is however clear that we will support the community to deliver their (meta)data according to the FAIR principles as much as possible and through the services offered in the project for the homeless data most of the data to be expected will flow through the already FAIR RI repositories and follow the FAIR workflow developed there.

All data and metadata will be published using persistent and unique identifiers such as ePIC and DOI. Final results will be in principle always be published through a Datacite DOI, in most cases of multiple files of different type as a data collection where each individual data object is identified by its own PID.

File naming conventions are irrelevant for FAIR data objects and each community and user can follow their own best practices when this is relevant for data use in the community or project.

Keywords follow the ENVRI-FAIR standard and will use for variable names the CF standard names (https://cfconventions.org/standard-names.html), where possible and appropriate. For keywords we recommend the GCMD nomenclature (https://earthdata.nasa.gov/earth-observation-data/find-data/idn/gcmd-keywords).

Raw data if curated by ATMO ACCESS will not be versioned but always consist of immutable data objects containing the data as generated now of calculation or measurement by the device, datalogger or instrument. Each data processing action in ATMO ACCESS will generate higher level data that is clearly annotated with metadata that provides the provenance information that at least explains, who (persons, affiliation), how (software, version), with which parameters (document or machine readable parameter file) and following which protocol (document, script or other machine readable workflow), all identified with persistent and unique identifiers and provided as data object with the dataset or linked through by open access. Each consecutive processing in the chain will deliver a next version of the dataset, the incremental version number will be part of the provenance metadata and consist of an indicator of major version changes followed by an indicator of minor changes, in the format of integer numbers separated by a decimal dot, such as for example "2.5".

We will ask the data providers that generate data in the TNA activities to follow the principles described here but cannot guarantee this. It might very well be the case that users that will make use of the TNA or VA homeless data service follow their own curation and intermediate storage and data processing and only want to store and make accessible final quality controlled data products through the project. In the case they make use of the homeless data service we will make sure that the (meta)data follows the respective FAIR data handling of the responsible RI data repository.

In that case metadata will follow the ENVRI-FAIR recommendations of the respective community and at least follow either Dublin Core, ISO19135 or DCAT AP v2.

2.2 Making data openly accessible:

- **Specify which data will be made openly available? If some data is kept closed provide rationale for doing so**
- **Specify how the data will be made available**
- **Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?**
- **Specify where the data and associated metadata, documentation and code are deposited**
- **Specify how access will be provided in case there are any restrictions**

In principal all data and software generated by the project will be open and be accessible through open licenses such a CC BY 4 and GPL 3.0, and shared through open repositories such as the RI repositories, Zenodo, Pangaea en Github.

Of course data from the TNA campaigns or data generated in the VA services can be part of other research projects and the data generators/owners will also have the need to publish the data under different licenses or keep part of the data under embargo until the research has been published. All RIs and repositories advised will support the possibility of embargo for the data and whenever possible offer support for other licenses than the default of the repository.

All data access through the RI and advised repositories already follow the FAIR principles and provide data through open and public protocols. Through the data curation of homeless data service we make sure that data formats are following the community open standards, in this case mostly NASA-AMES, ASCII and CF compliant netcdf and hdf.

2.3 Making data interoperable:

- **Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability.**
- **Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?**

Through the data curation of homeless data service we make sure that data formats are following the community open standards, in this case mostly NASA-AMES, ASCII and CF compliant netcdf and hdf.

Each data processing action in ATMO ACCESS will generate higher level data that is clearly annotated with metadata that provides the provenance information that at least explains, who (persons, affiliation), how (software, version), with which parameters (document or machine readable parameter file) and following which protocol (document, script or other machine readable workflow), all identified with persistent and unique identifiers and provided as data object with the dataset or linked through by open access.

2.4 Increase data re-use (through clarifying licenses):

- **Specify how the data will be licenced to permit the widest reuse possible**
- **Specify when the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed**
- **Specify whether the data produced and/or used in the project is useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why**
- **Describe data quality assurance processes**
- **Specify the length of time for which the data will remain re-usable**

In principal all data and software generated by the project will be open and be accessible through open licenses such a CC BY 4 and GPL 3.0, shared through open repositories such as the RI repositories, Zenodo, Pangaea en Github.

Of course data from the TNA campaigns or data generated in the VA services can be part of other research projects and the data generators/owners will also have the need to publish the data under different licenses or keep part of the data under embargo until the research has been published. All RIs and repositories advised will support the possibility of embargo for the data and whenever possible offer support for other licenses than the default of the repository.

The data will be kept for as long as the foreseen life-time of the RIs' and external repositories. All of these repositories have a long term perspective of at least 5 years from now and in most case an even longer time span (>20 years), with all providing contingency procedure to preserve the data outside the repository in the case the repository has to be terminated.

2.1.3. 3. Allocation of resources

Explain the allocation of resources, addressing the following issues:

- **Estimate the costs for making your data FAIR. Describe how you intend to cover these costs**
- **Clearly identify responsibilities for data management in your project**
- **Describe costs and potential value of long term preservation**

The costs of the FAIR provision of data and metadata is provided from the operational funding of the RIs. The marginal costs of the additonal storage of (meta)data from this project is provided by this project, that allows to develop the Virtual Access Services and support the operation of these during the second half of the project. Additional support for example of use of EOSC services like the B2SAFE replication of data is provided by other projects like EOSC eDICE and ENVRIFAIR.

The data management responsibilities are clearly defined in the project Description of Action and will be further detailed during the development of the services through for example the deliverables D5.1 and D5.3. Further details are described in the documents listed in section 2.6.

2.1.4. 4. Data security

Address data recovery as well as secure storage and transfer of sensitive data

All RIs provide secure storage of the submitted data and metadata and have in place redundancy of storage, with regular additional backups to restore the state of the repository in case of failures. In general the research data generated in this project is not sensitive or privacy relevant data.

The ICOS Carbon Portal repository is certified by CoreTrustSeal.

2.1.5. 5. Ethical aspects

To be covered in the context of the ethics review, ethics section of DoA and ethics deliverables. Include references and related technical aspects if not covered by the former

Ethics are in general not applicable on the research data generated in this field of research. All personal data will be processed according the European GDPR regulation and the applicable national laws for the different institutes.

2.1.6. 6. Other

Refer to other national/funder/sectorial/departmental procedures for data management that you are using (if any)

ICOS data management practices are described in: ICOS Data Policy and ICOS Data Lifecycle Document

ACTRIS data management practices are described in: https://github.com/actris/data-management-plan/blob/master/DMP/ACTRIS-DMP.md#

IAGOS data management data practices are described in: IAGOS Data Management Plan

NILU/EBAS data management practices are described in: https://ebas-submit.nilu.no