H2020 - Research and Innovation Action

APPLICATE

Advanced Prediction in Polar regions and beyond: Modelling, observing system design and LInkages associated with a Changing Arctic climaTE

Grant Agreement No: 727862

Deliverable No. 6.1

Data Management Plan

Version 1, August 2017

Submission of Deliverable

Work Package	WP6 Data and HPC Management		
Deliverable No	6.1		
Deliverable title	Data Management Plan		
Version	1		
Status	Draft		
Dissemination level	PU - Public		
Lead Beneficiary	MET Norway		
Contributors	x 1 – AWI	x 2 – BSC	☐ 3 - ECMWF
	☐ 4 – UiB	☐ 5 – UNI Research	☐ 6 – MET Norway
	☐ 7 – Met Office	☐ 8 – UCL	☐ 9 - UREAD
	☐ 10 – SU	☐ 11 – CNRS-GAME	☐ 12 - CERFACS
	☐ 13 – AP	☐ 14 – UiT	☐ 15 - IORAS
	☐ 16 - MGO		
Due Date	30 April 2017		
Delivery Date	Date		

This project has received funding from the European Union's Horizon 2020 Research & Innovation programme under grant agreement No. 727862.

Table of Contents

2 EXECUTIVE SUMMARY

This plan is based on the H2020 FAIR Data Management Plan (DMP) template designed to be applicable to any H2020 project that produces, collect or processes research data. This is the same plan as OpenAIRE is referring to in their guidance material. The purpose of the Data Management Plan is to describe the data that will be created and how, as well as the plans for sharing and preservation of the data generated. This plan is a living document that will be updated during the project.

APPLICATE follows a metadata-driven approach where a physically distributed number of data centres are integrated using standardised discovery metadata and interoperability interfaces for metadata and data. The APPLICATE Data portal, providing a unified search interface to all APPLICATE will also be able to host data. APPLICATE promotes free and open access to data in line with the European Open Research Data Pilot (OpenAIRE).

Within this plan an overview of the production chains for model simulations is provided as well as an initial outline of dissemination. This will be updated as the project progresses, with updates scheduled at project months 18 (April 2018), 36 (October 2019) and 48 (Octobenr 2020). The APPLICATE search interface for datasets will be available by August 2017.

3 INTRODUCTION

3.1 Background and motivation

The purpose of the data management plan is to document how the data generated by the project is handled during and after the project. It describes the basic principles for data management within the project. This includes standards and generation of discovery and use metadata, data sharing and preservation and life cycle management.

This document is a living document that will be updated during the project in time with the periodic reports (project months 18, 36 and 48).

APPLICATE 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).

3.2 Organisation of the plan

This plan is based on the H2020 FAIR Data Management Plan (DMP) template[1] designed to be applicable to any H2020 project that produces, collect or processes research data. This is the same plan as OpenAIRE is referring to in their guidance material.

4 Administration details

Project Name	APPLICATE
Funding	EU HORIZON 2020 Research and Innovation Programme
Partners	Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research (AWI) - Bremerhaven, Germany Barcelona Supercomputing Center - Barcelona, Spain European Centre for Medium-Range Weather Forecasts (ECMWF) - Reading, United Kingdom University of Bergen (UiB) - Bergen, Norway Uni Research AS - Bergen, Norway Norwegian Meteorological Institute (MET Norway) - Oslo, Norway Met Office - Exeter, United Kingdom Catholic University of Louvain (UCL) - Louvain-la-Neuve, Belgium The University of Reading (UREAD) - Reading, United Kingdom Stockholm University (SU) - Stockholm, Sweden National Center for Scientific Research (CNRS-GAME) - Paris, France (with contributions from Météo France) European Centre for Research and Advanced Training in Scientific Calculation

1 http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

	(CERFACS) - Toulose, France Arctic Portal - Akureyri, Iceland University of Tromsø (UiT) - Tromsø, Norway P.P. Shirshov Institute of Oceanology, Russian Academy of Sciences (IORAS) - Moscow, Russia Federal State Budgetary Institution Voeikov Main Geophysical Observatory (MGO) - St. Petersburg, Russia

5 Data summary

The overarching mission of APPLICATE is

To develop enhanced predictive capacity for weather and climate in the Arctic and beyond, and to determine the influence of Arctic climate change on Northern Hemisphere mid-latitudes, for the benefit of policy makers, businesses and society.

Therefore APPLICATE is primarily a project in which numerical models (for weather and climate prediction) are used. As such it depends on observations (e.g. for model evaluation and initialization), but the data generated by the project is primarily gridded output from the numerical simulations.

The APPLICATE data management system will be used to collect information of relevant third party datasets that the APPLICATE community could benefit from, and to share and preserve the datasets APPLICATE is generating, both internally and externally.

A full overview of the datasets to be generated is yet not fully known, but there is an overview of the production chains. This was prepared in the proposal and is provided in Tables 1–3 below.

5.1 Data overview

5.1.1 Types and formats of data generated/collected

APPLICATE will primarily generate gridded output resulting from numerical simulations and metrics based on these core datasets. The models used produce a number of output formats which is not known in detail, but specific requirements apply for data sharing and preservation (see below).

Self-explaining file formats (e.g. NetCDF, HDF/HDF5) combined with semantic and structural standards like the Climate and Forecast Convention will be used. The default format for APPLICATE datasets are NetCDF following the Climate and Forecast Convention (feature types grid, timeseries, profiles and trajectories if applicable). This includes the Coupled Model Intercomparison Project (CMIP) requirements. The NetCDF files must be created using the NetCDF Classic Model (i.e. compression is allowed, but not groups and compound data types). The ESGF CMOR is recommended for conversion of model output.

Some datasets may be made available as WMO GRIB or BUFR. Where no clear standard is identified initially, dedicated work will be attributed to identifying a common approach for those data.

APPLICATE will exploit existing data in the region. In particular operational meteorological data made available through WMO Gobal Telecommunication System will be important for the model experiments. No full overview of third party data that will be used is currently available. If necessary (required by the scientific community in APPLICATE) 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 APPLICATE scientists. There is however no plan initially to harvest the data.

Furthermore, model data produced in the context of CMIP5 and CMIP6 will be used as a baseline against which model improvements will be tested.

5.1.2 Origin of the data

Data will be generated by a suite of numerical models, including operational weather prediction and climate models. A preliminary list was provided in the proposal and is included below.

APPLICATE is primarily a project in which numerical models are used. As such it depends on observations (e.g. for model evaluation and initialization), but the data generated by the project is primarily gridded output from numerical simulations.

A summary of the numerical models to be used is provided in Tables 1–3.

Table 1: List of climate models.

Model	**AWI-CM**	**EC-Earth CNRM-CM**	**NorESM**	**HadGEM**
Partner	AWI	BSC, UCL, SU CNRS-GAME, CERFACS	UiB, UR, Met.no	MO, UREAD
Atmosphere	ECHAM6 T127 L95	IFS ARPEGE-Climat T255/T511 L91 T127/T359 L91	CAM-OSLO 1o×1o L32 / L46	MetUM N216/N96 L85
Ocean	FESOM Unstruct. mesh 15-100 km L41 4.5-80 km L41	NEMO NEMO 1o , 0.25o L75 1o, 0.25 o L75	NorESM-O (extended MICOM) 1o, 0.25o L75	NEMO 1o×1o L75 0.25o×0.25o L75
Sea ice	FESIM	LIM3 GELATO	CICE	CICE
Surface	JSBACH	HTESSEL SURFEX	SURFEX	JULES
CMIP6	Yes	Yes Yes	Yes	Yes

Table 2: List of subseasonal to seasonal prediction systems.

Model	**EC-Earth**	**CNRM-CM**	**IFS**	**HadGEM/GloSea**
Partner	BSC, UCL, AWI	CNRS-GAME	ECMWF	MO, UREAD
Atmosphere	IFS T255/T511 L91	ARPEGE Climat T255/T359 L91	IFS T511-T319 L91	MetUM N216 L85

Model	**EC-Earth**	**CNRM-CM**	**IFS**	**HadGEM/GloSea**
Ocean	NEMO 1°/0.25° L75	NEMO 1°/0.25°, L75	NEMO 1°, L75	NEMO 0.25o×0.25o L75
Sea ice	LIM3	GELATO	LIM2/3	CICE
Land	HTESSEL	SURFEX	HTESSEL	JULES
Data assimilation	Ensemble Kalman filter	Extended Kalman Filter SAM2	4D-Var	4D-Var, NEMOVAR 3D-Var FGAT

Table 3: Numerical weather prediction systems.

Model	**ARPEGE**	**AROME**	**IFS**	**AROME-Arctic**
Partner	CNRS-GAME	CNRS-GAME	ECMWF	Met.no
Atmosphere	ARPEGE T1198, stretched HR (7.5km on grid pole), L105	AROME 1.3km / 500m, 90 vertical levels	IFS T1279 L137	AROME 2.5 km L65
Ocean	N/A	N/A	N/A	N/A
Sea ice	GELATO	GELATO	N/A	SICE
Land	SURFEX	SURFEX	HTESSEL	SURFEX
Data assimilation	4D-Var	dynamical adaptation	4D-Var	3D-Var

At the moment of writing this data management plan, the total amount of data is not known. The information will be updates in further versions of the plan.

5.1.3 ECMWF YOPP data

Within APPLICATE, ECMWF has begun to generate an extended two-year global dataset to support the World Meteorological Organization's Year of Polar Prediction (YOPP). The start of production was timed to coincide with the official launch of YOPP in Geneva, Switzerland, on 15 May. The dataset is intended to support YOPP's goal of boosting polar forecasting capacity. In addition to the usual forecast data stored at ECMWF, it will include additional parameters for research purposes.

These include 'tendencies' in physical processes modelled in ECMWF's Integrated Forecasting System (IFS). More information on the ECMWF YOPP dataset is available from ECMWF. The actual data is available through the ECMWF YOPP Data Portal. This will be linked to the APPLICATE Data Portal.

5.2 Making data findable, including provisions for metadata [fair data]

APPLICATE is following a metadata driven approach, utilizing internationally accepted standards and protocols for documentation and exchange of discovery and use metadata. This ensures interoperability at the discovery level with international systems and frameworks, including WMO Information System (WIS), Year of Polar Prediction (YOPP), and many national and international Arctic and marine data centers (e.g. Svalbard Integrated Arctic Earth Observing System).

APPLICATE data management is distributed in nature, relying on a number of data centres with a long-term mandate. This ensures preservation of the scientific legacy. The approach chosen is in line with lessons learned from the International Polar Year, and the ongoing efforts by the combined SAON/IASC Arctic Data Committee to establish an Arctic data ecosystem.

APPLICATE promotes the implementation of Persistent Identifiers at each contributing data centre. Some have this in place, while others are in the process of establishing this. Although application of globally resolvable Persistent Identifiers (e.g. Digital Object Identifiers) is not required, it is promoted by the APPLICATE data management system. However, each contributing data centre has to support locally unique and persistent identifiers if Digital Object Identifiers or similar are not supported.

Concerning naming conventions, APPLICATE requires that controlled vocabularies are used both at the discovery level and the data level to describe the content. Discovery level metadata must identify the convention used and the convention has to be available in machine readable form (preferably through Simple Knowledge Organisation System). The fallback solution for controlled vocabularies is the Global Change Master Directory vocabularies.

The search model of the data management system is based on GCMD Science Keywords for parameter identification through discovery metadata. At the data level the Climate and Forecast Convention is used for all NetCDF files. For data encoded using WMO standards, GRIB and BUFR, the standard approach at the host institute is followed. All discovery metadata records are required to include GCMD Science Keywords. Furthermore, CMOR standards will be employed for some of the climate model simulations, especially those contributing to CMIP6.

Versioning of data is required for the data published in the data management system. Details on requirements for how to define a new version of a dataset is to be agreed, but the general principles include that a new version of a model dataset is defined if the physical basis for the model has changed (e.g. modification of spatial and temporal resolution, number of vertical levels and internal dynamics or physics). Integration of datasets (e.g. to create a long-time series) is encouraged, but these datasets must be clearly documented.

The APPLICATE data management system can consume and expose discovery metadata provided in GCMD DIF and ISO19115. If ISO19115 is used, GCMD keywords must be used to describe physical and dynamical parameters. Support for more formats is being considered. More specifications will be identified early in the project. As ISO19115 is a container that can be used in many contexts, APPLICATE promotes the application of the WMO Profile for discovery metadata. This is based on ISO19115. APPLICATE will be more pragmatic than WMO accepting records that not fully qualify in all aspects. The dialogue on what is required will be aligned with the ongoing efforts of the combined SAON /IASC Arctic Data Committee to ensure integration with relevant scientific communities.

APPLICATE will integrate with the YOPP Data Portal to make sure that APPLICATE datasets are discoverable through the YOPP Data Portal. This will be implemented letting the YOPP Data Portal harvest the relevant discovery metadata from the APPLICATE data catalogue.

5.3 Making data openly accessible [fair data]

All discovery metadata will be available through a web based search interface available through the central project website (applicate.met.no[2]). Some data may have temporal access restrictions (embargo period). These will be handled accordingly.

Valid reasons for an embargo period on data are primarily for educational reasons, allowing Ph.D. students to prepare and publish their work. Even if data constrained in the embargo period, data will be shared internally in the project. Any disagreements on access to data or misuse of data internally are to be settled by the APPLICATE Executive Board.

Data in the central repository will be made available through a THREDDS Data Server, activating OPeNDAP support for all datasets and OGC Web Map Service for visualisation of gridded datasets. Standardisation of data access interfaces and linkage to the Common Data Model through OPeNDAP[3] is promoted for all data centres contributing to APPLICATE. This enables direct access of data within analysis tools like Matlab, Excel[4] and R. Activation of these interfaces to data are recommended for other contributing data centres as well.

Metadata and data for the datasets are maintained by the responsible data centres (including the central data repository). Metadata supporting unified search is harvested and ingested in the central node (through applicate.met.no) where it will be made available through human (web interface) and machine interfaces (OAI-PMH, support for OpenSearch is considered).

Datasets with restrictions are initially handled by the responsible data centre. Generally the metadata will be searchable and contain information on how to request access to the dataset. An example of a dataset with access restrictions is the ECMWF YOPP dataset where user registration is required. Access to information about the dataset does however not require registration.

5.4 Making data interoperable [fair data]

In order to be able to reuse data, standardisation is important. This implies both standardisation of the encoding/documentation, as well as the interfaces to the data. Further up in the document, it is referred to documentation standards widely used by the modelling communities. This includes encoding model output as NetCDF files, following the Climate and Forecast convention or the WMO GRIB format. The WMN formats are table driven formats where the tables identifies the content and makes it interoperable. NetCDF files following the CF convention is self describing and interoperable. Application of the CF conventions implies requirements on the structure and semantic annotation of data (e.g. through identification of variables/parameters through CF standard names). Furthermore it requires encoding of missing values etc.

To simplify the process of accessing data, APPLICATE recommends all data centres to support OPeNDAP. OPeNDAP allows streaming of data and access without downloading the data as physical files. If OPeNDAP is not supported, straightforward HTTP access must be supported.

In order to ensure consistency between discovery level and use level metadata, a system for translation of discovery metadata keywords (i.e. GCMD Science keywords) to CF Standard

2 Will be established within August 2017.
3 http://apievangelist.com/2014/12/05/history-of-apis-noaa-apis-have-been-restful-for-over-20-years/
4 https://www.opendap.org/support/faq/general/use-spreadsheet

names is under development. This implies that e.g. controlled vocabularies used in the documentation of data may be mapped on the fly to vocabularies used by other communities. This is inline with current activities in the SAON/IASC Arctic Data Committee.

5.5 Increase data re-use (through clarifying licenses) [fair data]

APPLICATE promotes free and open data sharing in line with the Open Research Data Pilot. Each dataset needs a license attached. The recommendation in APPLICATE is to use Creative Commons attribution license for data. See https://creativecommons.org/licenses/by/3.0/ for details.

APPLICATE data should be delivered in a timely manner meaning without un-due delay. Any delay, due or un-due, shall not be longer than one year after the dataset is finished. Discovery metadata shall be delivered immediately.

APPLICATE is promoting free and open access to data. Some data may have constraints (e.g. on access or dissemination) and may be available to members only initially. Furthermore, some of the data will be used for modelling development purposes and are thus of limited interest to the broader community; these data will not be made publicly available. A draft dissemination plan was outlined in the proposal and is provided in Table 4. This will be updated as the project progresses.

Table 4: Draft data dissemination plan.

Purpose	**Model systems**	**Experimental design**	**Data**
Determine the impact of model enhancements on process representation and systematic model error (WP2)	• AWI-CM • EC-Earth • CNRM-CM • NorESM • HadGEM	Baseline data: CMIP6-DECK experiments Implement the model changes suggested in WP2 in coupled models: • 200-yr pre-industrial control experiments • CMIP6 historical experiments • 1% CO 2 increase experiments	Partial Dissemination
Determine Arctic-lower latitude linkages in atmosphere and ocean (WP3)	Coupled models • AWI-CM • EC-Earth • CNRM-CM • NorESM • HadGEM	Large ensembles (50-100 members) of 12-months experiments starting June 1st with sea ice constrained to observed and projected sea ice fields Multi-decadal experiments with and without artificially reduced Arctic sea ice (enhanced downwelling LW radiation over sea ice); use of tracers for the ocean Repeat with enhanced models	Full Dissemination

Purpose	Model systems	Experimental design	Data
	Atmospheric models • ECHAM6 • IFS • ARPEGE-Climat • CAM-OSLO • MetUM	Large ensembles (50-100 members) of 12-months experiments starting June 1st with sea ice constrained to observed and projected sea ice fields Various corresponding sensitivity experiments to explore the role of the background flow, and the prescribed sea ice pattern Repeat with enhanced models	Full Dissemination
	Seasonal prediction systems • EC-Earth • CNRM-CM	Seasonal prediction experiments with and without relaxation of the Arctic atmosphere towards ERA-Interim reanalysis data: 9-member ensemble forecasts with members initialized on Nov 1st, Feb 1st, May 1st and Aug 1st for the years 1979-2016 and 1993-2016 for EC-Earth and CNRM-CM, respectively.	Full Dissemination
Arctic observing system development (WP4)	Atmospheric model • IFS	Data denial experiments with the IFS for key observations (snow, surface pressure, wind, moisture) and different seasons.	Partial dissemination
	Seasonal prediction • EC-Earth • HadGEM • GloSea	- Perfect model experiments to characterize basic sensitivity of forecasts to initial conditions. - Different configurations of initial conditions using reanalyses, new observations, ocean reruns forced by atmospheric reanalyses. - Experiments focused on sea-ice thickness, snow and spatial data sampling	Partial dissemination
Determine the impact of APPLICATE model enhancements on weather and climate prediction (WP5)	Atmospheric model • ARPEGE • AROME • IFS • AROME-Arctic	Test recommendations for model enhancements made in WP2 in pre-operational configurations Explore the impact of nesting, driving model and resolution	Partial dissemination

Purpose	Model systems	Experimental design	Data
	Seasonal prediction • EC-Earth • CNRM-CM • HadGEM	Test recommendations for model enhancements made in WP2 in pre-operational configurations	Partial dissemination
	Climate change • AWI-CM • EC-Earth • NorESM • AWI-CM	Establish the impact of model enhancements developed in WP2 on climate sensitivity by carrying out experiments using the same initial conditions and time period (1950—2050) employed in HiResMIP climate sensitivity by carrying out experiments using the same initial 2050) employed in HiResMIP climate sensitivity by carrying out experiments using the same initial conditions and time period (1950—2050) employed in HiResMIP	Partial dissemination

The quality of each dataset is the responsibility of the Principal Investigator. The Data Management System will ensure the quality of the discovery metadata and that datasets are delivered according to the format specifications.

Numerical simulations and analysed products will be preserved for at least 10 years after publication.

6 Allocation of resources

In the current situation it is not possible to estimate the cost for making APPLICATE data FAIR. Part of the reason is that this work is relying on existing functionality at the contributing data centres and that this functionality has been developed over years. The cost of preparing the data in accordance with the specifications and initial sharing is covered by the project. Maintenance of this over time is covered by the business models of the data centres.

A preliminary list of data centres involved is given in Table 5.

Table 5: Preliminary list of data centres involved in the APPLICATE project. A full list will be provided in the next update to the plan (due April 2018).

Data centre	URL	Contact	Comment
DKRZ	http://www.dkrz.de	Thomas Jung	TBC
Norwegian Meteorological Institute/Arctic Data Centre	https://applicate.met.no/	Øystein Godøy	This subsystem will provide a unified search interface to all the data APPLICATE is generating. It will also host data not being hosted by other data centres contributing to APPLICATE. Metadata interfaces are available, data interoperability supported using OGC WMS and OPeNDAP. Will integrate relevant data from WMO GTS.
BSC	https://www.bsc.es	Pierre-Antoine Bretonnière	

Each data centre is responsible for accepting, managing, sharing and preserving the relevant datasets. Concerning interoperability interfaces the following interfaces are required:

1. Metadata
    - OAI-PMH serving either CCMD DIF or the ISO19115 minimum profile with GCMD Science Keywords. Dedicated sets should be available to identify APPLICATE data in large data collections.
2. Data (will also use whatever is available and deliver this in original form, for those data no synthesis products are possible without an extensive effort)
    - OGC WMS (actual visual representation, not data)
    - OPeNDAP

In the current situation there is no overview of the costs of long-term preservation of data as this is the responsibility of the contributing data centres and the business model for these differs. This information will be updated in further versions of the DMP.

All data that will contribute to CMIP6 will be stored in data centres contributing to the Earth System Grid Federation (ESGF). APPLICATE data centres contributing to this will be shown in the table above. For APPLICATE, the experiments contributing to the Polar Amplification Model Intercomparison Project (PA-MIP) will be managed in a ESGF data centre.

7 Data security

Data security relies on the existing mechanisms of the contributing data centres. APPLICATE recommends ensuring the communication between data centres and users with secure HTTP. Concerning the internal security of the data centre, APPLICATE recommends the best practises from OAIS. The technical solution will vary between data centres, but most data centres have solutions using automated check sums and replication.

The central node relies on secure HTTP, but not all contributing data centres support this yet.

8 Ethical aspects

APPLICATE is not concerned with ethical sensitive data and follows the guidance of the IASC Statement of Principles and Practises for Arctic Data Management.

9 Other

APPLICATE is linked to WMO's Year of Polar Prediction activity. In this context APPLICATE is relating to the WMO principles for data management identified through the WMO Information System.