esiwace
CENTRE OF EXCELLENCE IN SIMULATION OF WEATHER AND CLIMATE IN EUROPE

# Annex: Data Management Plan

Deliverable 5.4

# Version 1.0, Feb. 2016

This project has received funding from the European Union's
Horizon 2020 research and innovation programme
under grant agreement No 675191

## About this document

**Work package in charge:** WP5 Management & Dissemination
**Actual delivery for this deliverable:** 28 February 2016
**Dissemination level**: PU (for public use)

**Lead author:** German Climate Computing Center (DKRZ), Project office, Kerstin Fieg, Chiara Bearzotti
**Other contributing partners:**
German Climate Computing Center (DKRZ), Julian Kunkel
European Centre for Medium-Range Weather Forecasts (ECMWF), Daniel Thiemert, Peter Bauer
Centre National de Recherche Scientifique - Institut Pierre-Simon Laplace (CNRS-IPSL), Sylvie Joussaume
Barcelona Supercomputing Center (BSC), Oriol Mula-Valls, Kim Serradell
Max Planck Institute for Meteorology (MPI-M), Reinhard Budich
Centre Européen de Recherche et de Formation Avancée en Calcul Scientifique (CERFACS), Sophie Valcke
Science and Technology Facilities Council (STFC), Martin Juckes
Sveriges Meteorologiska och Hydrologiska Institut (SMHI), Uwe Fladrich

**Contacts:** esiwace@dkrz.de
**Visit us on**: www.esiwace.eu
**Follow us on Twitter:** @esiwace

## Index

# 1. Executive Summary

The Data Management Plan (DMP) of ESiWACE gives an overview of available research data, access and the data management and terms of use. The DMP reflects the current state of the discussions, plans and ambitions of the ESiWACE partners, and will be updated as work progresses.

# 2. Introduction

### Why a Data Management Plan (DMP)?

It is a well-known phenomenon that the amount of data is increasing while the use and re-use of data to derive new scientific findings is more or less stable. This does not imply, that the data currently unused are useless - they can be of great value in future. The prerequisite for meaningful use, re-use or recombination of data is that they are well documented according to accepted and trusted standards. Those standards form a key pillar of science because they enable the recognition of suitable data.
To ensure this, agreements on standards, quality level and sharing practices have to be negotiated. Strategies have to be fixed to preserve and store the data over a defined period of time in order to ensure their availability and re-usability after the end of ESiWACE

### What kind of data are considered in the DMP?

The main purpose of a Data Management Plan (DMP) is to describe *Research Data* with the metadata attached to make them *discoverable*, *accessible*, *assessable*, *usable beyond the original purpose* and *exchangeable* between researchers.
According to the "Guidelines on Open Access to Scientific Publication and Research Data in Horizon 2020" (2015)*:*

> "*Research data* refers to information, in particular facts or numbers, collected to be examined and considered and as a basis for reasoning, discussion, or calculation. In a research context, examples of data include statistics, results of experiments, measurements, observations resulting from fieldwork, survey results, interview recordings and images. The focus is on research data that is available in digital form."

However, the overall objective of ESiWACE is to improve efficiency and productivity of numerical weather and climate simulations on HPC systems by enhancing the scalability of numerical models, foster the usability of community wide used tools and pursue the exploitability of model output.
Thus ESiWACE focuses more on the production process and tools than on production of research or observation data and so the amount of *Research Data* which ESiWACE intend to produce is limited, at least at this stage of the project.

### What can be expected from ESiWACE DMP?

In the following we will describe the lifecycle, responsibilities and review processes and data management policies of research data, produced in ESiWACE. The DMP reflects the current status of discussion within the consortium about the data that will be produced. It is not a fixed document, but evolves during the lifespan of the project.
The target audience of the DMP is all project members and research institutions using the data and data produced.

# 3. Register on numerical data sets generated or collected in ESiWACE

The register has to be understood as living document, which will be updated regularly during project lifetime. The intention of the DMP is to describe numerical model or observation datasets collected or created by ESiWACE during the runtime of the project.
The information listed below reflects the conception and design of the individual work packages at the beginning of the project. Because the operational phase of the project started in January 2016, there is no dataset generated or collected until delivery date of this DMP.

The data register will deliver information according to Annex 1 of the Horizon 2020 guidelines (2015) (in *italics)*:

- **Data set reference and name:** *Identifier for the data set to be produced.*
- **Data set description:** *Descriptions of the data that will be generated or collected, its origin (in case it is collected), nature and scale and to whom it could be useful, and whether it underpins a scientific publication. Information on the existence (or not) of similar data and the possibilities for integration and reuse.*
- **Standards and metadata***: Reference to existing suitable standards of the discipline. If these do not exist, an outline on how and what metadata will be created.*
- **Data sharing***: Description of how data will be shared, including access procedures, embargo periods (if any), outlines of technical mechanisms for dissemination and necessary software and other tools for enabling re-use, and definition of whether access will be widely open or restricted to specific groups. Identification of the repository where data will be stored, if already existing and identified, indicating in particular the type of repository (institutional, standard repository for the discipline, etc.).*
  *In case the dataset cannot be shared, the reasons for this should be mentioned (e.g. ethical, rules of personal data, intellectual property, commercial, privacy-related, security-related).*
- **Archiving and preservation (including storage and backup)**: *Description of the procedures that will be put in place for long-term preservation of the data. Indication of how long the data should be preserved, what is its approximated end volume, what the associated costs are and how these are planned to be covered*

### 3.1 Datasets collected within WP1

| **WP1 Governance, Engagement and long-term sustainability** | |
|---|---|
| **What types of data will the project generate/collect?** | WP1 is not going to generate numerical data sets. |

### 3.2 Datasets collected within WP2

| **WP2 Scalability** | |
|---|---|
| **Data set reference and name** | EC-Earth model output and performance data |
| **Data set description** | EC-Earth high-resolution model output will be generated for test runs. Furthermore, performance data will be collected. |

| | |
|---|---|
| | Constraints: IFS data may not be used for commercial purpose. |
| **Standards and metadata** | Model output will be in NetCDF and GRIB.<br><br>No metadata is automatically generated by the model. CMIP6-compliant metadata generation may become available during the course of the project.<br><br>No quality check is applied automatically. If necessary, CMIP6 compliant quality checking may be applied. |
| **Data Sharing** | EC-Earth model data and performance data will be shared (if useful):<br>- Within the ESiWACE project, particularly WP2<br>- Within the EC-Earth consortium<br>- Within the ENES community, particularly the IS-ENES2 project<br><br>Data sharing will generally be through access to the HPC systems or data transfer to shared platforms.<br><br>If common experiments are run in the context of other projects (e.g. PRIMAVERA, CMIP6), data publication may be through ESGF. |
| **Archiving and preservation (including storage and backup)** | Long-term data storage will most likely not be needed for the data created in this project, the exception being potential common experiments with other projects. In the latter case, data storage will be provide by the respective projects. |
| **Reported by** | Uwe Fladrich (uwe.fladrich@smhi.se) |

| | |
|---|---|
| **Data set reference and name** | BSC Performance Analysis |
| **Data set description** | In WP2, BSC will carry on performance analysis and modifications to the source code of the earth system models to run in others programming models (like OmpSs).<br>While the modified model code is no data to be described here, the performance analysis will produce trace outputs that contain the information of an execution of the model. In this case, the size can be a constraint. On many-core systems, the traces generated by a complex model can have a very big size (more than hundreds of gigabytes) so this can be a problem to share this information between partners. The integration and the reuse of this information would not be a problem if the different actors take a first decision in the tools to be used in these performance analyses. |
| **Standards and metadata** | All the tools to trace executions provide information about the format of the outputs and how to read them. Moreover, some of these tools can convert formats to improve the compatibility.<br>Data can be in a raw binary or text format. In this last case, CSV or XML are usual formats to deal with the information.<br>In the case of Paraver tool, in each trace there is a file describing which events are in the trace. This file usually contains a code and a text description for each event. |

| | |
|---|---|
| **Data Sharing** | For the traces, a repository allowing the distribution of big files must be implemented. If the distribution is individual and sporadic, a solution like an FTP can fit to the requirement. If we want to setup a repository with all the traces for further analyses, another solution must be deployed. The solution will have to classify data among the model run, the platform, the configuration. This can lead to a big number of different combinations. |
| **Archiving and preservation (including storage and backup)** | Codes will be stored in the gitlab, during the time that the partners consider it convenient, but for the traces, due to the high volume of the data generated, another strategy has to be designed. Long term storage solution (like tapes) could be a good solution. Traces are usually a collection of big files suited to be stored in tape solution archive. |
| **Reported by** | Kim Serradell (kim.serradell@bsc.es) |

| | |
|---|---|
| **Data set reference and name** | IFS and OpenIFS model output. |
| **Data set description** | IFS and OpenIFS model integrations will be run and standard meteorological and computing performance data output will be generated. Both will be run at ECMWF, and only performance data will be made available to the public. The meteorological output will be archived in MARS, as it is standard research experiment output. The data will be used for establishing research and test code developments, and will enter project reports and generally accessible publications.<br>The IFS will not be made available, OpenIFS is available through a dedicated license. |
| **Standards and metadata** | IFS meteorological output (incl. metadata) and format follows WMO standards.<br>Compute performance (benchmark) output will be stored and documented separately. Data will be in ASCII and maintained locally. The output will be reviewed internally, and the ECMWF facilities allow reproduction of this output if necessary. |
| **Data Sharing** | All output can be shared within the ESiWACE consortium, and is primarily located in the ECMWF archiving system MARS.<br>Data provision to the public is limited for meteorological output, and it adheres to the ECMWF data policy. Access can be granted in individual cases.<br>Computing performance output can be made publicly available. This output can be managed by the ESiWACE website. |
| **Archiving and preservation (including storage and backup)** | As no large quantities of data will be produced, there are no requirements for long-term data management. The experiment output is stored in MARS that is backed up regularly.<br>Volumes and cost are negligible. |
| **Reported by:** | Peter Bauer (peter.bauer@ecmwf.int) |

| | |
|---|---|
| **Data set reference and name** | |
| **Data set description** | WP2 will extend the benchmark suite fro coupling technologies |

| | |
|---|---|
| | currently developed in IS-ENES2 to target new platforms with O(10K-100K) cores accessible during the ESiWACE longer timeframe. OASIS, OpenPALM, ESMF, XIOS and YAC will be considered.<br>Benchmark suites for I/O libraries and servers will have to be built from scratch. The inter- comparison will include XIOS, ESMF and CDI-pio.<br>A subset of the results of these benchmarks for specific technologies on specific computing platforms will be collected and made available as a reference. |
| **Standards and metadata** | The data per se will be just text files containing numbers (e.g. the communication time for a specific coupling exchange as a function of the number of cores used to run the coupled components) and will not adhere to any specific standard.<br>The metadata attached to the data will contain the revision number of the benchmark sources that will be managed under SVN or GIT and a description of the parameters tested for a specific set of results (e.g. number of cores, number of coupling fields, etc.). The metadata will appear also as a text file (in the form of a Readme file) available in the data directory.<br>The results of the benchmarks will be reviewed by the participating IS-ENES2 partners and reported in ESiWACE D2.1 |
| **Data Sharing** | The benchmark sources (managed under SVN or GIT) and subset of results will be freely accessible to all. The description on how to access the sources and results will be available on ESiWACE web site. |
| **Archiving and preservation (including storage and backup)** | The subset of benchmark results and associated metadata will be uploaded to a data centre (e.g. DKRZ) and attached with a standard data DOI. Specific subset of results data will curated and preserved as a reference to compare with for the people who would want to run the benchmark themselves for O(10) years and will be regularly replaced by new subsets of new tests for new platforms. |
| **Reported by:** | Sophie Valcke (sophie.valcke@cerfacs.fr) |

### 3.3 Datasets collected within WP3

| **WP3 Usability** | |
|---|---|
| **What types of data will the project generate/collect?** | WP3 is not going to generate typical numerical data sets, WP3 is going to produce papers and reports, and to some extent software code. |

### 3.4 Datasets collected within WP4

| **WP4 (Exploitability)** | |
|---|---|

| | |
|---|---|
| **What types of data will the project generate/collect?** | WP4 (Task 4.3) will generate semantic mappings between metadata standards. The mappings will be made available through a SPARQL server and curated at STFC and ECMWF |

### 3.5 Datasets collected within WP5

| **WP5 Management and Dissemination** | |
|---|---|
| **What types of data will the project generate/collect?** | WP5 is not going to generate numerical data sets |

## 4. References (*Bibliography*)

Guidelines on Data Management in Horizon 2020, Version 2.0, 30 October 2015:
http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf
Guidelines on Open Access to Scientific Publication and Research Data in Horizon 2020, Version 2.0, 30 October 2015
http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

## 5. Glossary

| | |
|---|---|
| DOI | Digital Object Identifier |
| DMP | Data Management Plan |
| EC | European Commission |
| GRIB | GRIdded Binary format, WHO |
| H2020 | Horizon 2020, EU funding Strategy for 2014 - 2020 |
| pdf | Portable Document Format |
| NetCDF | NETwork Common Data Format |
| ppt | Power Point |
| RDF | Resource Description Framework |
| SPARQL | **S**PARQL **P**rotocol **a**nd **R**DF **Q**uery **L**anguage |