H2020-JTI-EuroHPC-2019-1

Project no. 956748

ADAPTIVE MULTI-TIER INTELLIGENT DATA MANAGER FOR EXASCALE

D1.2
Data Management Plan

Version 1.0

Date: October 17, 2021

Type: Deliverable
WP number: WP1

Editor: Jesus Carretero
Institution: UC3M

Project co-funded by the European Union Horizon 2020 JTI-EuroHPC research and innovation programme and Spain, Germany, France, Italy, Poland, and Sweden		
Dissemination Level		
PU	Public	√
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Change Log

Rev.	Date	Who	Site	What
1	21/04/21	Jesus Carretero	UC3M	Document creation.
2	17/08/21	Jesus Carretero	UC3M	DMP definition
3	06/09/21	Javier Garcia-Blas	UC3M	DMP repositories definition
4	14/09/21	Emmanuel Jeannot	INRIA	Deliverable review
5	15/09/21	Massimo Torquati	CINI	Deliverable review

Executive Summary

Data management is a crucial issue for research and innovation projects and many mistakes were made in the past, when no one was actually thinking about what to do with the data and how to preserve or make them available for other researchers too. This first Data Management Plan (DMP) shows that there are mainly data sets that will be produced as part of the project activities and that are relevant to be included in the DMP. The DMP describes the types of data that will be generated or gathered during the project, the standards that will be used, the ways how the data will be exploited and shared for verification or reuse, and how the data will be preserved.

The DMP is a living document, which will evolve during the lifespan of the project, particularly whenever significant changes arise such as dataset updates or changes in Consortium policies. This document is the first version of the DMP, delivered in Month 6 of the project. It includes an overview of the datasets to be produced by the project, and the specific conditions that are attached to them. Although this report already covers a broad range of aspects related to the **ADMIRE** data management, the upcoming versions will get into more detail on particular issues such as data interoperability and practical data management procedures implemented by the **ADMIRE** project consortium.

This document has been produced following these guidelines and aims to provide a consolidated plan for **ADMIRE** partners in the DMP policy that the project will follow.

Contents

Chapter 1

Introduction

The purpose of the Data Management Plan (DMP) deliverable is to provide relevant information concerning the data that will be collected and used by the partners of the **ADMIRE** project.

The DMP describes the data management life cycle for all datasets to be collected, processed and/or generated by the research project. It covers:

- How data should be handled during and after the project.
- What types and formats of data will be generated/collected.
- Which methodologies and standards will be applied.
- Whether the data be shared or made open-access, and how
- How data will be curated and preserved during the project as well as after its conclusion.

This is the first version of the DMP. It contains preliminary information about the data generated by the project, whether and how it will be exploited or made accessible for verification and re-use, and how it will be curated and preserved. The purpose of the Data Management Plan is to provide an analysis of the main elements of the data management policy that the consortium will use regarding all datasets that the project will generate. The DMP is a live document, thus it will evolve during the lifespan of the project. By following the EC "Guidelines on FAIR data management in Horizon 2020" [4], the DMP will be updated over the course of the project whenever significant changes arise such as new data, changes of consortium policies, or of consortium composition.

The European Union enables *Open Innovation* by requiring that projects funded under the European Union Framework Programme for Research and Innovation, Horizon 2020, must ensure open access (free of charge, online access for any user) to all peer-reviewed scientific publications relating to the project's results.

The DMP specifies the implementation of the pilot, in particular concerning the data generated and collected, the standards in use, the workflow to make the data accessible for use and verification by the community and define the strategy of curation and preservation of the data. Thus, we refer to the **ADMIRE** Grant Agreement (GA), Articles 29.1 and 29.2 about project data dissemination, as reported in the following:

> Regarding the digital research data generated in the action ("data"), the beneficiaries must:
>
> a) Deposit in a research data repository, as specified in Chapter 3 and take measures to make it possible for third parties to access, mine, exploit, reproduce and disseminate free of charge for any user the following:
>
> (i) the data, including associated metadata, needed to validate the results presented in scientific publications as soon as possible;
>
> (ii) other data, including associated metadata, as specified and within the deadlines laid down in the 'data management plan' (see Annex A);

b) Provide information via the repository about tools and instruments at the disposal of the beneficiaries and necessary for validating the results (and whenever possible to provide the tools and instruments themselves).

The remaining part of the document is organized as follows:

- Chapter 2 describes the general principles used to organize project data
- Chapter 3 describes the data repositories used through the project
- Annex A describes the deadlines for data publication

Chapter 2

General Principles

This chapter describes the main aspects to be considered in ADMIRE's data management, that must be followed by all partners of the project. In general terms, ADMIRE's research data should be 'FAIR', that is findable, accessible, interoperable and re-usable. [4].

2.1 Data sources

There are three major sources of data in the frame of the project, as shown below.

Software codes (SM) : According to the IPRs (open sources or proprietary), the codes will be organized in a GitLab repository hosted by UC3M.

Project deliverables (PD) : Support and intermediate documents, scientific papers and technical reports (ST), media materials including posters, videos, photos, and meeting notes (MM) will be associated with the public website or wiki system, according to their access rules (public or confidential).

Test and validation data (DS) : They will be linked to the public website or wiki system, according to their access rules (public or confidential).

Every document or data file generated must have a number. The document editor/coordinator will be responsible of increasing the version number using a version control system.

2.2 Primary research data generated by the project

The project's primary research data will be coming from performance figures, figures supporting software engineering metrics, and other results used to support the research publications produced during the project lifespan. The research is expected to yield results in terms of, e.g., parallel performance (execution time and speedup), reduction in errors, improved time to solution and improved maintainability.

Full consideration will be given to allowing proper statistical analysis of the results, using means, standard deviations, regression tests, and other relevant metrics. Performance data and software engineering results will be derived from parallel source programs. Where there are no copyright or commercial confidentiality issues, these sources will also be released.

Raw data resulting from research is the data that has not been coded, grouped, refined or modified in any way. Even if raw data has more potential usages than modified data, as a general rule, the ADMIRE project will not provide raw data in open repositories, due to data quality checking and IPR controls. Instead, each partner should keep her sets of raw data, which should be mantained untouched, if possible.

2.3 Confidentiality and data protection concerns

To increase dissemination and data re-use, the majority of the data described above will be non-confidential, will not refer to human subjects, and will not introduce any security concerns. However, all research data will be collected and stored in line with European legislation on data protection, as relevant.

For software and data, open-source licenses are promoted in the project. While some partners could retain part of the results with other licenses, this will not be the norm. This rule is explained in the project's Consortium Agreement. Thus, data be licensed to permit the most extensive re-use possible as soon as possible in the project timeline. Embargoes are not foreseen for software and data.

Data will remain available after the end of the project in open data portals such as Zenodo [5] and eArchive [2], assuming that no budget is needed to keep this data open and reusable.

2.4 Preferred data formats

To simplify processing and avoid possible problems with transcribing data between evolving data formats, data will generally be stored as plain text.

For research data exchange, CSV format is preferred. However, a description of the data will be included with each dataset stored in the repository.

For documents and other dissemination data of the ADMIRE project, templates are provided. Those templates are mandatory to enhance compatibility and inspection inside the project.

All data elements must incorporate attribution of their original source, date and authors. If several contributions are made along the time, a changelog is required.

Table 2.1 shows the formats preferred and also those that are accepted in the ADMIRE data catalog. Preferred formats have been chosen because they are suggested to have the highest probability of maintaining accessibility and readability in the future. Accepted formats are commonly used formats that have good prospects of remaining readable in the long term.

Document type	**Preferred format**	**Accepted format**
Text documents	plain (.txt)	MS Word
	PDF	Rich Text Format
	Tex	
Markup language	JSON	XML
		XML
Spreadsheets	CSV	MS Excel (.xls)
		OpenDocument Spreadsheet
Images	PNG	TIFF
	SVG	JPEG, PDF
Databases	CSV	SQL
	JSON	OpenDocument Base

Table 2.1: Suggested data formats in ADMIRE

2.5 Tools for validating results

Complete information will be provided in the research publications about the tools that have been used to produce the results, including details of operating system versions, libraries and specific software tools, as relevant.

The majority of the software tools that will be used in the project will be free, open-source software, either produced by third parties or produced by the project and disseminated under open licenses. Full care will, however, be taken to avoid releasing proprietary software and to achieve the best possible commercial exploitation for tools that are developed and/or modified in the course of the project.

Where tools are restricted, it will generally be possible to obtain them under some sufficiently liberal license agreement to allow full reproduction of the research experiments (perhaps on payment of a fee).

2.6 Data security

All research data underpinning publications will be made available for verification and re-use unless there are justified reasons for keeping specific datasets confidential. The main elements when considering confidentiality of datasets are:

- Protection of intellectual property regarding new processes, products, and technologies where the data could be used to derive sensitive information that would impact the competitive advantage of the consortium or its members,
- Commercial agreements as part of the procurement of components or materials that might foresee the confidentiality of data
- Personal data that might have been collected in the project where sharing them is not allowed by the national and European legislation.

2.7 Data quality

Data quality control will be part of the project's quality management plan. Inaccurate data is challenging to detect, but some procedures to check them have been put in place in ADMIRE project to remove as many errors as possible before data publication:

- Double review of deliverables.
- Quality supervision at the work package level of the data produced in each work package.
- Peer review of documents, software, and papers.

Documentation and deliverables will be hierarchically organized via a knowledge base whose elements will be readable by all project partners. Write access, administration, and use of quality-control tools will be available for those responsible.

UC3M, as project coordinator, will be ultimately responsible for the data quality control and of coordinating data quality plan. However, the data quality control will be organized hierarchically in a bottom-up process:

- Data producers will be primarily responsible for ensuring data quality and conformance with the required standards.
- Peer reviewers will be in charge of data quality reviews and co-responsible of potential issues not reported.
- Work package leaders will be responsible for controlling the data quality of the data produced in their WPs.

- Scientific Coordinator will be the next control level for data produced as result of research.
- Project coordinator will be ultimately responsible for the data quality control and of coordinating data quality plan.

Each data contributor should take reasonable steps to check the accuracy of the data and report any errors both for data and metadata, even if they will be found after publication.

2.8 Source code

ADMIRE's source code will be programmed using the coding standard defined in the project. This coding standard is not mandatory for already-established projects that are already using their own coding standard, such as GekkoFS, that could keep on using the existing coding standard.

2.9 Allocation of resources

Resources for data management plan (DMP) are mostly allocated to the Project Management work package (WP1) and Project Dissemination and Exploitation work package (WP8).

Project partners have their own budget foreseen for publication in Open Access journals or public repositories.

2.10 Ethics

ADMIRE project does not foresee to handle data with any ethical or legal issues that can have an impact on data sharing.

Chapter 3

Data Repositories

This chapter describes the repositories to be used in the ADMIRE project for public and private data management.

3.1 Public repositories

ADMIRE project will publish data mainly in open access repositories to increase the impact of the project an to promote results exploitation. Open-access repositories, such as an institutional repository or disciplinary repository, provide free access to research for users outside the institutional community and are one of the recommended ways to achieve the open access vision described in the Budapest Open Access Initiative [1] definition of open access.

For storing, preserving, and sharing data, ADMIRE will rely on standard repositories that will provide services for managing data in a secure and persistent manner.

For software development, we will set up internal repositories (e.g., GitLab), while for software distribution we will rely on the Gitlab of the project and standard repositories such as GitHub and Bitbucket.

For other types of data, ADMIRE will leverage the European data management infrastructures such as OpenAIRE [6] and EUDAT [3]portals, as most project partners maintain OpenAIRE-compatible electronic repositories and others can freely use Zenodo [5], a repository hosted by CERN.

Subject to the provisions above and with due concern for commercial issues, all the research data will be made available through Zenodo [5]. This repository has been funded by the European Union with the specific objective of storing the research data that will be generated by Framework 7 and Horizon 2020 research projects. Using this repository has several advantages over the self-hosting or use of an institutional repository that we had initially envisaged.

- It provides a more visible and centralized location for accessing research data, covering many EU projects and other sources of research data, rather than a single project and institution. This improves data dissemination.
- Data in the repository can be accessed free of charge. A large number of standard licenses are supported.
- Data in the repository can easily be searched, mined and otherwise exploited.
- The portal automatically generates a DOI and metadata for each dataset that is submitted. The DOI provides a location-independent reference point that is intended to ensure long-term accessibility of the data. The DOI can be referred to directly from within research publications.
- Because it is dedicated to this purpose, the repository provides a long-term storage solution, ensuring the longevity and relevance of the research data. In particular, generated data is guaranteed to be stored and curated long after the project has finished. Without continual intervention, this would be difficult to achieve using local and/or institutional repositories.

Moreover, we will store project research data results on EUDAT Research Data Management (RDM) platform [3], in particular the data sets derived from experimental evaluations of project technologies are good candidates for its B2Share store. BSC is one of the EUDAT partners and will help the project to choose and understand the adequate data management services to make the data produced in the project interoperable and to allow data exchange and re-use between researchers, institutions, and organizations in different countries.

The metadata that is recorded for each dataset will include the date of submission, the owner of the data, a description of the data content, and a link to the **ADMIRE** project.

3.1.1 eArchive

The project website will be maintained after the end of the project. However, to ensure long-term continuity and value, data will be transferred to an institutional data repository.

To that purpose we will use the institutional repository from UC3M (e-archivo.uc3m.es) which is federated with several international open platforms (OpenAIRE, OpenDOAR, EUDAT, BASE, and CORE).

3.2 Consortium private repository

The project consortium employs a private repository for storing all the information necessary to conduct the project. This includes deliverables, journal and conference publications, internal presentations, and media.

The repository is based on *GitLab* and it is located at `https://gitlab.arcos.inf.uc3m.es:8380/admire/project.git`. This site is maintained and supported by the project coordinator and fullfils the security rules imposed by Universidad Carlos III of Madrid:

- The site supports encryption by using SSL certificates.
- Only authorized users can access inside the repository.
- This site provides mechanisms for brute-force attack to protect non-authorized accesses.

UC3M, as project coordinator institution, can manage individual and team access to the organization's repositories. All members of the project were granted access to the private repository at the beginning of the project. New requests are granted on-demand to the project coordinator.

Appendix A

Data Management Plan deadlines

Data from the project should be made available in the repositories respecting the following deadlines:

- Public reports should be available in the project Web page and the eArchive repository at maximum one month after their publication in the EU project portal.
- . Open access scientific papers references should be cited in the Web page at maximum one month after their acceptance. A PDF copy should be made also available at the ResearchGate project portal (https://www.researchgate.net/project/ADMIRE-4).
- For non-open access publications, a PDF preprint copy should be made also available at the ResearchGate project portal (https://www.researchgate.net/project/ADMIRE-4) at maximum one month after their acceptance. PDF of the acceptted version should also be made available after the embargo period.
- The data, including associated metadata, needed to validate the results presented in scientific publications as soon as possible.
- other data, including associated metadata, should be made available in Zenodo and EUDAT at maximum six months after their creation.

Bibliography

[1] Budapest open access initiative. Technical report, Budapest, Hungary. https://www.budapestopenaccessinitiative.org, 2002.

[2] Universidad Carlos III de Madrid. Archivo abierto institucional de la universidad carlos iii de madrid. Technical report, Universidad Carlos III de Madrid, Madrid. Spain. https://e-archivo.uc3m.es, 2021.

[3] EUDAT. Store and publish your research data. Technical report, EUDAT Collaborative Data Infrastructure., Finland. https://b2share.eudat.eu, 2020.

[4] Directorate-General for Research & Innovation EUROPEAN COMMISSION. Guidelines on fair data management in horizon 2020. Technical report, EUROPEAN COMMISSION., Brussels, Belgium. https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf, 2016.

[5] European Organization For Nuclear Research and OpenAIRE. Zenodo. Technical report, CERN, Switzerland. https://www.zenodo.org/, 2013.

[6] OpenAIRE. Open science in europe. Technical report, European Union's Horizon 2020 Grant Agreement No. 777541. https://www.openaire.eu/, 2021.