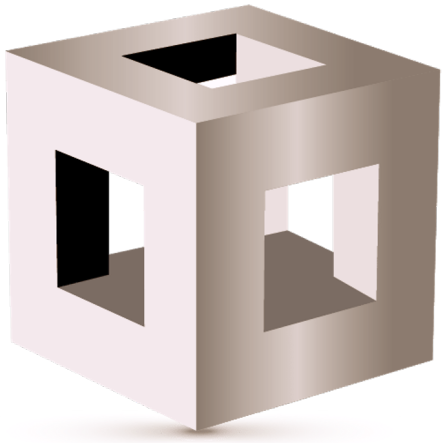
HORIZON Research and Innovation Actions
HORIZON-CL4-2022-DIGITAL-EMERGING-01
European Health and Digital Executive Agency



OpenCUBE - Open-Source Cloud-Based Services on EPI Systems

GA 101092984

D 1.1

Data Management Plan

WP1: Management

Funded by
the European Union

Date of preparation (latest version): 30/06/2023

DOCUMENT INFORMATION

Deliverable Number	D1.1
Deliverable Name	Data Management Plan
Due Date	30/06/2023
Deliverable Lead	KTH
Authors	Ivy Peng (KTH), Stefano Markidis (KTH)
Responsible Author	Ivy Peng (KTH), ivybopeng@kth.se
Keywords	Information Governance, Lifecycle Management, Validation and Quality Assurance, Access Control Policies, Information Security Measures, Resource Compliance
WP/Task	WP 1 / Task 1.2
Nature	DMP
Dissemination Level	PU
Final Version Date	30/06/2023
Reviewed by	Utz-Uwe Haus (HPE), Celine Scetbun (SiPearl)

DOCUMENT HISTORY

Partner	Date	Comment	Version
KTH	6/11/2023	Skeleton version of the deliverable	0.1
KTH	6/25/2023	Updated version of the deliverable after internal review	0.2
KTH	6/28/2028	Final edit before submission	1.0

Executive Summary

Data management is crucial for research and innovation projects for governing data organization, data integrity and quality, data sharing and collaboration. A data management plan (DMP) ensures that data is effectively managed throughout the project lifecycle, promotes good research practices, and enhances the usability, integrity, and long-term value of the project's data.

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. The DMP describes the types of data that will be generated or gathered during the project, the standards that will be used, the ways the data will be exploited and shared for verification or reuse, and how the data will be preserved.

The DMP is a living document that will evolve during the lifespan of the project, particularly whenever significant changes arise, such as dataset updates or changes in Consortium policies. Although this report already covers a broad range of aspects related to the OpenCUBE data management, the upcoming versions will go into more detail on issues such as data interoperability and practical data management procedures implemented by the OpenCUBE project consortium.

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

Contents

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 OpenCUBE project.

The DMP describes the data management life cycle for all datasets to be collected, processed, and 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 that will be 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 living document. Thus, it will evolve during the lifespan of the project. The DMP will be updated over the course of the project whenever significant changes arise, such as new data, changes in 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 Europe, 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 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 defines the strategy of curation and preservation of the data. Thus, we refer to the OpenCUBE Grant Agreement (GA) about project data dissemination, as reported in the following:

1.2.9 Open Science Practices, Open Research data management, and management of other research output

The output of the OpenCUBE project will be a blueprint of a validated European Cloud computing software stack. That means that the design documents, the evaluation reports, and the software modified or developed for the prototype system will be made available under liberal Open Source and Open Data licenses. All journal or conference publications will be available as *Green* open-access, allowing for major dissemination of the scientific outcomes of the project.

We do not expect large-scale research data to be generated in the project, but benchmarking data may be part of the research reports or scientific papers and will be provided in accordance with FAIR Guiding Principles for scientific data management and stewardship through a suitable Open Science repository or the publisher with Green Open Access, as appropriate.
Code contributed to existing open-source projects will be donated to the respective project under the existing licenses. The new code will be licensed under a liberal license, e.g., BSD-3-clause.

The publications, code artifacts, and design documents generated by OpenCUBE will be published through the dissemination channels described in Section 2.2, particularly 2.3.3.

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 throughout the project.
- Chapter 4 describes the initial data research output plan for the project.
- Chapter 5 presents a brief timeline of the data management plan.
- Chapter 6 concludes the deliverable.

2 General principles

This chapter describes the main aspects to be considered in OpenCUBE's data management that all partners of the project must follow. In general terms, OpenCUBE's research data should be 'FAIR,' which is findable, accessible, interoperable, and reusable.

2.1 FAIR research data management

OpenCUBE will create a common platform for accessing performance data, analysis, and white papers generated by the project. The performance data will respect the FAIR principles (findable, accessible, interoperable, and reusable).

OpenCUBE will produce performance data from running benchmarks and applications on the software stack used within the OpenCUBE project. Performance data will consist of execution logs (trace files) as well as profile data, such as hardware performance counter samples. This performance data is the basis for analysis that will be based on a variety of tools and methods. The data generated with the help of these tools will be homogenized in the form of written performance reports that will be made public. These reports will always include artifacts that enable the reproducibility of the results, such as the software build setup, the hardware specification of the experiment environment, the input parameters, the input datasets, the execution parameters of the simulation, the result correctness check data, and the recorded performance data. The artifact will also point to the exact version (`release` or `git-hash`).

2.2 Data sources

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

1. **Software codes (SM)**: According to the IPRs (open sources or proprietary), the codes developed in the OpenCUBE project will be organized in a common GitHub repository.

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

3. **Test and validation data (DS)** will be hosted in a common repository. Each data set will be assigned an identifiable ID. They will be linked to the public website 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 for increasing the version number using a version control system.

2.3 Primary data generated by the project

The project's primary data will be coming from benchmarking results, performance figures, figures supporting system deployment and software engineering metrics, and other results used to support the research publications produced during the project lifespan.

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 software programs. Where there are no copyright or commercial confidentiality issues, these sources will also be released.

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

2.4 Confidentiality and data protection concerns

To increase dissemination and data reuse, most of the data described above will be non-confidential, not refer to human subjects, and not introduce any security concerns. Non-disclosure of information, such as the preliminary benchmark results based on SiPearl and SemiDynamics samples, is defined in the project's Consortium Agreement Section 10. 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, assuming that no budget is needed to keep this data open and reusable.

2.5 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 OpenCUBE 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 over time, a changelog is required.

Table 2.1 shows the preferred formats and those accepted in the OpenCUBE 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
	YAML	XML
Spreadsheets	CSV	MS Excel (.xls)
		OpenDocument Spreadsheet
Images	PNG	TIFF
	SVG	JPEG, PDF
Databases	CSV	SQL
	JSON	OpenDocument Base
		n-triple

Table 2.1: Suggested data formats in OpenCUBE

2.6 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. Most 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 during the project.

2.7 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 the 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 national and European legislation.

2.8 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 the OpenCUBE 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. The data quality control will be organized hierarchically in a bottom-up process:

- Data producers will primarily ensure data quality and conformance with the required standards.
- Peer reviewers will oversee data quality reviews and be co-responsible for 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 a result of the research.
- The project coordinator will be ultimately responsible for the data quality control and coordinating the 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.

2.9 Allocation of resources

Resources for the data management plan (DMP) are mostly allocated to Project Management in the work package (WP1) and Project Dissemination and Exploitation in the work package (WP6). Project partners have their own budget foreseen for publication in Open Access journals or public repositories.

2.10 Ethics

The OpenCUBE project does not foresee handling data with any ethical or legal issues that can have an impact on data sharing. Should such data arise, the coordinator needs

to be notified immediately, and this data management plan will be amended appropriately.

3 Data repositories and management of Intellectual Properties

This chapter describes the repositories to be used in the OpenCUBE project for public and private data management and management of intellectual properties.

3.1 Public repositories

The OpenCUBE project will publish data mainly in open-access repositories to increase the impact of the project and to promote results exploitation. Open-access repositories, such as institutional repositories or disciplinary repositories, provide free access to research for users outside the institutional community.

For storing, preserving, and sharing data, OpenCUBE will rely on standard repositories on the project's GitHub organization https://github.com/opencube-horizon and the project's website for sharing public data.

For software development, we will use a common GitHub organization https://github.com/opencube-horizon for hosting public repositories for software distribution.

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.

Basecamp [https://3.basecamp.com/4072478/projects/26765975] repositories (for the whole project and each WP) are set up for hosting project documents and essential management tools. Basecamp also provides functionalities corresponding to mailing lists, and they are being used to keep members informed.

KTH, as the 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 by the project coordinator.

3.3 Management of Intellectual Properties

With respect to Intellectual Property Rights (IPR), all backgrounds owned by the partners and the IPR policies have been clearly stated in the project's Consortium Agreement Section 9 Access Rights and Attachment 1. IP created during the project will be the property of the partner who creates it. However, all IP created during the project will be available to other partners for use on the project without payment. Use of that property following the conclusion of the project will be subject to normal considerations.

The software developed during the project and scientific and technical publications will be offered in open access unless an explicit mention in the GA prohibits its distribution to protect IP. Several open-source software will be further developed and used in OpenCUBE to support the deployment of cloud services on EPI systems. The table below summarizes the licenses and open-source repositories used in the project.

Library / Tool	Use	Licence	Repository
Kubernetes	an open source system for managing containerized applications across multiple hosts	Apache License Version 2.0	https://github.com/kubernetes/kubernetes
SUSE	A Linux Operating System	Multiple licenses available	https://github.com/SUSE/kernel
CSM	the HPE Cray System Management software	the MIT Open Source licence	https://github.com/Cray-HPE/csm
OpenFAM	Fabric-attached memory	BSD 3-Clause License	https://github.com/OpenFAM/OpenFAM
Volcano	A cloud native batch system for Kubernetes	Apache License Version 2.0	https://github.com/volcano-sh/volcano
iPIC3D	an open-source C++ and MPI Particle-in-Cell code for the simulation of space plasma	Apache License Version 2.0	https://github.com/KTH-HPC/iPIC3D
AutoDock-GPU	open-source docking engines	GNU GENERAL PUBLIC LICENSE Version 2	https://github.com/ccsb-scripps/AutoDock-GPU
Apache Spark	a unified analytics engine for large-scale data processing	Apache License Version 2.0	https://github.com/apache/spark
FDB	a domain-specific object store developed at ECMWF for storing,	Apache License Version 2.0	https://github.com/ecmwf/fdb

	indexing and retrieving GRIB data		
Mamba	A software for Managed Abstracted Memory Array	BSD-3 license	https://gitlab.com/cerl/mamba

4 Data Research Outputs

The purpose of data research outputs is to communicate, disseminate, and share the findings and insights derived from research data with the scientific community, policymakers, practitioners, and the public in the European Union.

The table below shows our initial data research output plan for OpenCUBE.

Types of data/ research outputs	**Findability of data/ research outputs**	**Accessibility of data/ research outputs**	**Interoperability of data/ research outputs**	**Reusability of data/ research outputs**	**Costs and responsible team**
Design specification, Performance, Benchmark Data	Open-Access reports and white papers	Immediate	Pdf files, trace files, text files with configuration data and tabular output	Documents available for analysis and reproducibility studies	No Cost / All consortium
Software tools, Software documentation	Open-source on GitHub	OpenCUBE source codes will be made available under different licensing schemes	C, C++, Fortran, Python, IDL source codes	Codes available for deployment	No Cost / All consortium

5 Data Management Plan Timeline

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

- Public reports should be available on the project webpage and repository for a maximum of one month after their publication in the EU project portal.
- Open access scientific papers references should be cited in the project webpage a maximum of one month after their acceptance.
- The data, including associated metadata, is needed to validate the results presented in scientific publications as soon as possible.

6 Conclusion

This deliverable presented the OpenCUBE DMP, which ensures that data is effectively managed throughout the project lifecycle, promotes good research practices, and enhances the project's data's usability, integrity, and long-term value. This deliverable includes an overview of the datasets to be produced by the project and the specific conditions that are attached to them.

This document is the first version of the DMP, delivered in Month 6 of the project. The DMP is a living document. Thus, it will evolve during the lifespan of the project. The DMP will be updated over the course of the project whenever significant changes arise, such as new data, changes in consortium policies, or of consortium composition.