



HORIZON EUROPE FRAMEWORK PROGRAMME

CloudSkin

(grant agreement No 101092646)

Adaptive virtualization for AI-enabled Cloud-edge Continuum

D2.2 Data Management Plan, 1st version

Due date of deliverable: 30-06-2023
Actual submission date: 29-06-2023

Start date of project: 01-01-2023

Duration: 36 months

Summary of the document

Document Type	Data Management Plan
Dissemination level	Public
State	v1.0
Number of pages	11
WP/Task related to this document	WP1 / T1.1
WP/Task responsible	Universitat Rovira i Virgili (URV)
Leader	Marc Sanchez-Artigas (URV)
Technical Manager	Marc Sanchez-Artigas (URV)
Quality Manager	Julen Bohoyo (URV)
Author(s)	Vanesa Ruana (URV)
Partner(s) Contributing	URV
Document ID	CloudSkin_D2.2_Public.pdf
Abstract	First version of the Data Management Plan. Experiments, workloads, benchmarks, and results will be delivered as open data to the community. This deliverable will evolve during the lifetime of the project in order to reflect the current status of data management.
Keywords	Data Management Plan, Open Access, Open Research Data, FAIR data, ORDP

History of changes

Version	Date	Author	Summary of changes
0.1	13-04-2023	Vanesa Ruana	First draft.
0.2	19-05-2023	Several partners	Add the datasets.
1.0	29-06-2023	Vanesa Ruana	Final version.

Table of Contents

1	Executive summary	2
2	Data Summary	3
3	FAIR data	6
3.1	Making data findable	6
3.2	Making data openly accessible	7
3.3	Making data interoperable	7
3.4	Increase data re-use (through clarifying licenses)	7
3.5	Management principles	7
4	Allocation of resources	8
5	Data security	8
6	Ethical aspects	9
7	Data Management Plan review process and timetable	9
8	Conclusions	10

List of Abbreviations and Acronyms

API	Application Programming Interface
CC	Creative Commons
Cholec80	Endoscopic video dataset containing 80 videos of cholecystectomy surgeries from University of Strasbourg
CSV	Comma-separated values
DMP	Data Management Plan
DOI	Digital Object Identifier
DSAD	Dresden Surgical Anatomy Dataset
HeiChole	Surgical Dataset for surgical workflow and skill analysis from University Hospital Heidelberg und NCT

1 Executive summary

CloudSkin is committed to good data management. In an effort to provide a management life-cycle of the data needed to validate results in scientific publications, this first version of the Data Management Plan (DMP) has been provided as deliverable D2.2. This DMP describes how the research data was made findable, accessible, interoperable and reusable. This version also presents a summary of the existing datasets that are currently known to have been used over the course of the project.

As the project progresses and data is identified and collected, further information about generated data will be provided.

2 Data Summary

CloudSkin project wants to enable open access and reuse of the research data generated by Horizon Europe projects. CloudSkin has the commitment to:

- Develop (and keep up-to-date) a Data Management Plan (DMP).
- Deposit the project's data in a research data repository.
- Ensure third parties can freely access, mine, exploit, reproduce and disseminate our data.
- Provide related information and identify (or provide) the tools needed to use the raw data to validate our research.

In particular, the project applies to:

- The data (and associated metadata) needed to validate the results presented in scientific publications.
- Other curated and/or raw data (and associated metadata) that is specified within this Data Management Plan.

The main goal of the CloudSkin project is to design a cognitive cloud continuum platform to fully exploit the available Cloud-edge heterogeneous resources, finding the "sweet spot" between the cloud and the edge, and smartly adapting to changes in application behavior via AI. To facilitate automatic deployment, mobility and security of services, CloudSkin will build an innovative universal container-like execution abstraction based on WebAssembly that allows the seamless and trustworthy execution of (legacy) applications across the Cloud-edge continuum.

Table 1 presents a summary of the existing datasets that will be processed to validate the results of the CloudSkin project.

Table 1: Used datasets

UD1	
Name:	Dresden Surgical Anatomy Dataset (DSAD) [1]
Origin:	University Hospital Dresden and National Center for Tumor Diseases Dresden
Access:	Dresden Surgical Anatomy Dataset
Volume:	11 different organs, 33 patients (20.5G)
Variety:	Organ Segmentation
Frequency of update:	NA
UD2	
Name:	HeiChole [2]
Origin:	University Hospital Heidelberg and National Center for Tumor Diseases Dresden
Access:	HeiChole Dataset
Volume:	33 patients, ~200GB
Variety:	Surgical Workflow Analysis
Frequency of update:	NA

UD3	
Name:	Cholec80 [3]
Origin:	University of Strasbourg
Access:	Cholec80
Volume:	80 patients, ~80GB
Variety:	Surgical Workflow Analysis
Frequency of update:	NA
UD4	
Name:	Synthetic video data [3]
Origin:	National Center for Tumor Diseases Dresden
Access:	Synthetic Video Data
Volume:	~30GB
Variety:	Surgical Workflow Analysis and Segmentation
Frequency of update:	NA
UD5	
Name:	Agricultural data analysis - Campo de Cartagena (WDC)
Origin:	WIDHOC Smart Solutions
Access:	Agricultural technical laboratory
Volume:	13 sensors from nov 2021 to June 2023. 15,6 MB
Variety:	Aggregated data from agricultural sensors
Frequency of update:	30 mins
UD6	
Name:	AI-Enabled Optimization of a Machine Learning Service in METASPACE
Origin:	METASPACE
Access:	Deposited on our GitHub: https://github.com/metaspac2020/offsample
Volume:	23,238 ion images from 87 public spatial metabolomics datasets
Frequency of update:	NA
UD7	
Name:	AI-Enabled Optimization of Molecular Annotation in METASPACE
Origin:	METASPACE
Access:	Public datasets from multiple labs, academic institutes and a pharma company (AstraZeneca), deposited on our GitHub: https://github.com/metaspac2020/Lithops-METASPACE
Volume:	16 GB in total
Frequency of update:	NA

Aside from these datasets and benchmarks, the CloudSkin project will likely generate other data to validate the results presented in scientific publications (test data, APIs, source code used to perform analysis, etc.). All this data will be made available as open data and its re-use will be encour-

aged. As the project progresses and data is identified and collected, further information on data details will be provided. Table 2 presents a summary of the already generated datasets in the process of validating the results of the CloudSkin project.

Table 2: Generated dataset

GD1	
Name:	Surgical Phantom Liver Data (currently not public)
Description:	Continuously recorded phantom data from the Experimental Operating Room at the NCT
Access:	Data will be made accessible to project partners up request
Volume:	Increasing (currently ~ 20GB)
Variety:	
DOI:	NA

CloudSkin data will not only be useful for the current and future generation of big data and cloud technologies researchers, but also big data practitioners and companies (from SMEs to multinationals) with a vested interest in new programming models for data analytics.

3 FAIR data

In general terms, research data should be **FAIR**, that is **findable, accessible, interoperable and reusable** [4].

3.1 Making data findable

Used data In order to ensure that the data used in the project is easily findable, we will make an effort to include standard identification mechanisms in all our publications, source code and tutorials. Although not all datasets used in the project provide these identification mechanisms, we will take special care to provide the necessary instructions, metadata and tools for locating and processing those datasets.

Produced data CloudSkin is expected to deposit generated data in an open online research data repository. We have selected Zenodo as our data repository of choice. Zenodo is an OpenAIRE and CERN collaboration that allows researchers to deposit both publications and data, providing tools to link related items through persistent identifiers and data citations. Zenodo automatically assigns a Digital Object Identifier (DOI) to each item to make them easily and uniquely citable. Moreover, Zenodo is set up to facilitate the finding, accessing, re-using and interoperating of data sets, which are the basic principles that ORD projects must comply with.

To this end, we have created a CloudSkin community in Zenodo¹ to gather all the open data contributions of the project. The repository allows to assign specific keywords to each dataset as well as a minimum of the DataCite's Metadata Schema [5] recommended terms.

Whenever possible (according to publisher copyright policies regarding open access), research publications will also be uploaded to this repository to ensure the maximum dissemination of the results of the project. Publications will be linked to its associated research data.

Source code. To make the source code open to the general public, we have created a code repository in GitHub for CloudSkin². GitHub is currently one of the most popular code management systems due to the advanced features and easy management that it provides to developers. This has various potential benefits to the management and dissemination of CloudSkin source code: for instance, GitHub is well-known across developer communities, which facilitates the access to the source code of CloudSkin. Moreover, GitHub offers a plenty of options to fork/branch/merge versions of a software project that enables third-parties to easily extend the source code developed in CloudSkin (even for internal use). Additionally, we'll also make source code citable and uniquely identifiable by automatically archiving software releases in Zenodo [6].

¹<https://zenodo.org/communities/cloudskin-eu/>

²<http://github.com/cloudskin-eu>

As of the last release of this document, the CloudSkin Github profile contains 1 individual repository hosting software results.

Finally, the CloudSkin web page³ will list all project results and provide links to their respective repositories in Zenodo or GitHub.

3.2 Making data openly accessible

It is our intention that all data produced during the CloudSkin project is openly accessible as the default. Pre-existing datasets used in the experiments are mostly public and openly available (see Table 1).

Potential users will find out about the data through publications and the CloudSkin website. Data will be made available on publication of the associated paper and will be made accessible through the Zenodo repository.

3.3 Making data interoperable

Interoperability of data produced within the CloudSkin project is promoted through best practices. Data formats should adhere to widely used standards and should be compliant with available software applications. Where possible, standard codes will be followed (e.g.: ISO 639 for language codes, ISO 3166 for country codes, NUTS for region codes, ...).

As the project progresses and data is identified and collected, further information on making data interoperable will be outlined in subsequent versions of the DMP. Specifically, information on data and metadata vocabularies, standards or methodology to follow to facilitate interoperability and whether the project uses standard vocabulary for all data types present to allow interdisciplinary interoperability.

3.4 Increase data re-use (through clarifying licenses)

Data will be made accessible, and therefore available for re-use, within one month of the publication of the related peer-reviewed scientific article. Data will be shared under the Creative Commons Attribution 4.0 International Public License (CC BY 4.0) [7]. This license guarantees the widest possible re-use and redistribution while only requiring that appropriate credit is given.

The shared data will remain re-usable after the end of the project by anyone interested in it, with no access or time restrictions.

As the project progresses and data is identified and collected, further information on making data re-usable will be outlined in subsequent versions of the DMP. In specific, information about data quality assurance processes.

3.5 Management principles

The protocol below summarizes the management principles behind making generated research data FAIR:

³<http://cloudskin.eu>

PROTOCOL: Storing generated research data in CloudSkin project and making it FAIR

Beneficiaries will follow these procedures for each dataset collected or generated during the CloudSkin project:

- Store and make findable the dataset in the CloudSkin community of the Zenodo repository.
- Ensure that publications and research data behind them are cross-referencing each other through standard identification mechanisms.
- Ensure that each dataset provides metadata, particularly regarding access rights, licenses, and funding information.
- Each Work Package Leader is responsible for storing relevant research data to the repository.
- Data will be made accessible within one month of the publication of the related peer-reviewed scientific article.

Beneficiaries will follow these procedures for source code generated during the CloudSkin project:

- Store the source code under the CloudSkin organization in GitHub repository.
- Provide a comprehensive README file with instructions to run the code.
- Store each release of the source code to Zenodo repository and cross-reference related datasets and publications.
- Each Work Package Leader is responsible for storing relevant source code to the repository.

4 Allocation of resources

Regarding Open Access to research data, archiving at Zenodo is free of charge. Storing source code at the GitHub repository is also free of charge. Therefore, no costs are currently foreseen regarding the long term preservation of data.

The project coordinator has the ultimate responsibility for the data management in the project.

5 Data security

As CloudSkin delegates the archiving of data to Zenodo, their policies regarding data security apply:

- **Replicas:** All data files are stored in CERN Data Centres, primarily Geneva, with replicas in Budapest. Data files are kept in multiple replicas in a distributed file system, which is backed up to tape on a nightly basis.
- **Retention period:** Items will be retained for the lifetime of the repository. This is currently the lifetime of the host laboratory CERN, which currently has an experimental programme defined for the next 20 years at least.
- **File preservation:** Data files and metadata are backed up nightly and replicated into multiple copies in the online system.
- **Fixity and authenticity:** All data files are stored along with a MD5 checksum of the file content. Files are regularly checked against their checksums to assure that file content remains constant.
- **Succession plans:** In case of closure of the repository, best efforts will be made by CERN to integrate all content into suitable alternative institutional and/or subject based repositories.

6 Ethical aspects

There is no sensitive ethical issue of collecting, storing, processing and archiving data raised by the research of the CloudSkin project. Any potential ethical issue raised during the life of the project may be reported to the CloudSkin project board, which would, if necessary, raise immediate awareness of internal consortium members' executives, in order to take appropriate actions to resolve this issue.

Concerning potential ethical conflicts all issues will be resolved through the procedures depicted in relative legal documents (e.g., Consortium Agreement) and Commission guidelines.

7 Data Management Plan review process and timetable

As a *living* document, the Data Management Plan will be updated periodically. Particularly, the DMP will be updated whenever significant changes arise, such as:

1. New data
2. Changes in consortium policies (e.g. new innovation potential, decision to file for a patent)
3. Changes in consortium composition and other external factors (e.g. new member joining or current member leaving)

An up-to-date version will be available in time with each periodic review of the project. Table 3 summarizes the scheduled updates of the Data Management Plan.

Table 3: Timetable for Data Management Plan updates

Deliverable title	Del. No.	Month	Date
Data Management Plan, 1st version	D2.2	M6	June 2023
Data Management Plan, 2nd version	D2.4	M18	June 2024
Data Management Plan, 3rd version	D2.6	M36	December 2025

8 Conclusions

This document is the first version of the CloudSkin Data Management Plan. Its objective is to provide a clearly defined management strategy of data for the consortium. The CloudSkin DMP explains the life-cycle for the project data, whether it is collected, processed, or generated by the consortium. This DMP describes how the research data has been, and will be made findable, accessible, interoperable and reusable.

References

- [1] M. Carstens, F. M. Rinner, S. Bodenstedt, A. C. Jenke, J. Weitz, M. Distler, S. Speidel, and F. R. Kolbinger, "The dresden surgical anatomy dataset for abdominal organ segmentation in surgical data science," Scientific Data, vol. 10, no. 1, pp. 1–8, 2023.
- [2] M. Wagner, B.-P. Müller-Stich, A. Kisilenko, D. Tran, P. Heger, L. Mündermann, D. M. Lubotsky, B. Müller, T. Davitashvili, M. Capek, et al., "Comparative validation of machine learning algorithms for surgical workflow and skill analysis with the heichole benchmark," Medical Image Analysis, vol. 86, p. 102770, 2023.
- [3] A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. De Mathelin, and N. Padoy, "Endonet: a deep architecture for recognition tasks on laparoscopic videos," IEEE transactions on medical imaging, vol. 36, no. 1, pp. 86–97, 2016.
- [4] M. Wilkinson and et al, "The FAIR Guiding Principles for scientific data management and stewardship," Nature Scientific Data, no. 160018, 2016.
- [5] DataCite Metadata Working Group, "DataCite Metadata Schema for the Publication and Citation of Research Data. Version 4.2. DataCite e.V.." <https://doi.org/10.5438/bmjt-bx77>, 2019.
- [6] GitHub, "Making your code citable." <https://guides.github.com/activities/citable-code/>, 2016.
- [7] Creative Commons, "Creative Commons Attribution 4.0 International Public License." <https://creativecommons.org/licenses/by/4.0/legalcode>, 2013.