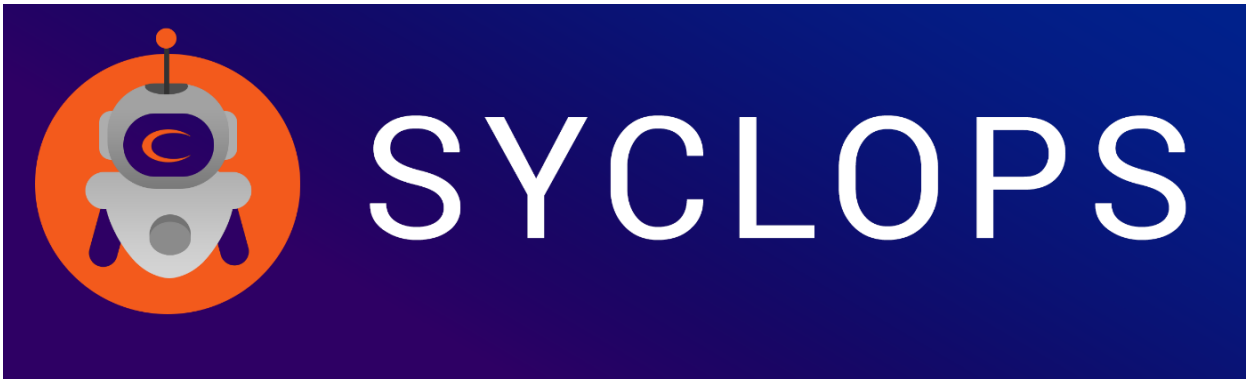
Deliverable 1.2 – Data Management Plan

GRANT AGREEMENT NUMBER: 101092877



Project acronym:	**SYCLOPS**
Project full title:	**Scaling extreme analYtics with Cross architecture acceLeration based on OPen Standards**
Call identifier:	**HORIZON-CL4-2022-DATA-01-05**
Type of action:	**RIA**
Start date:	**01/01/2023**
End date:	**31/12/2025**
Grant agreement no:	**101092877**

D1.2 – Data Management Plan

Executive Summary:	D1.2 describes mechanisms to handle data created or generated during the project.		
WP:	WP1 Project Management		
Author(s):	Kumudha Narasimhan, Mehdi Goli, Uwe Dolinsky, Raja Appuswamy		
Leading Partner:	CPLAY		
Participating Partners:	All Partners		
Version:	1.0	**Status:**	Draft
Deliverable Type:	R-Document	**Dissemination Level:**	PU – Public
Official Submission Date:	30-06-2023	**Actual Submission Date:**	05-08-2023

Disclaimer

The SYCLOPS consortium consists of the following partners:

No.	Partner Organisation Name	Partner Organisation Short Name	Country
1	EURECOM	EUR	**FR**
2	INESC ID - INSTITUTO DE ENGENHARIADE SISTEMAS E COMPUTADORES, INVESTIGACAO E DESENVOLVIMENTO EM LISBOA	INESC	**PT**
3	RUPRECHT-KARLS-UNIVERSITAET HEIDELBERG	UHEI	**DE**
4	ORGANISATION EUROPEENNE POUR LA RECHERCHE NUCLEAIRE	CERN	**CH**
5	HIRO MICRODATACENTERS B.V.	HIRO	NL
6	ACCELOM	ACC	FR
7	CODASIP S R O	CSIP	CZ
8	CODEPLAY SOFTWARE LIMITED	CPLAY	UK

Document Revision History

Version	Description	Contributions
0.1	25/07/2023 – 1st draft of the deliverable	All partners
1.0	31/7/2023 – Final draft	EUR

Authors

Author	Partner
Kumudha Narasimhan	CPLAY
Mehdi Goli	CPLAY
Uwe Dolinsky	CPLAY
Raja Appuswamy	EUR

Reviewers

Name	Organisation
Aleksandar Ilic	INESC
Vincent Heuveline	UHEI
Axel Naumann	CERN
Nimisha Chaturvedi	ACC
Pavel Zaykov	CSIP
Fred Buining	HIRO

Table of Contents

Executive Summary

The document provides the initial data management plan of the SYCLOPS project. The deliverable aims to define a framework outlining the SYCLOPS policies for data management, sharing, and protection during and after the duration of the project covering topics such as data, metadata content and format, policies for access, sharing and reuse, as well as long-term storage. During the project, the Data Management Plan will be continually assessed and re-evaluated to discover if it has been affected by future results of the work performed in all technical Work Packages. Therefore, the initial framework presented in this deliverable will further evolve during the project lifetime as a living document.

1 Introduction

The document provides an initial Data Management Plan (DMP) concerning the data processed, generated, and preserved during and after the SYCLOPS project. In addition, any topics of discussion regarding data usage, ethics, and security are also discussed in this deliverable. In a nutshell, deliverable D1.2 establishes a framework for the data management policy of the SYCLOPS project.

Towards this objective, the data, metadata, code, content, and format, sharing policies, storage, and personal data protection measures (if applicable) are entailed. This deliverable will be continually assessed during the project and updated accordingly.

1.1 Background

Deliverable D1.2 - DMP is part of Work Package (WP1) "Project Management" and reports on the activities concerning Task T1.3 covering the time period from the beginning of the project. It is the first version of the DMP.

1.2 Purpose and Scope

This deliverable defines a data management framework for the SYCLOPS project and addresses the following questions:

- What types of data will the Action generate/collect?
- What standards will be used?
- How will this data be exploited and/or shared/made accessible for verification and reuse?
- How will this data be curated and preserved?

1.3 Document structure

The structure of the document is as follows:

- Section 1 provides the deliverable's background, purpose and scope, giving its overall structure.
- Section 2 describes the SYCLOPS project strategy for Data collection.
- Section 3 details the Documentation and Metadata strategy.
- Section 4 comments on the ethical aspects to be considered in SYCLOPS during the use of the data.
- Section 5 entails the storage and backup of the data generated.
- Section 6 details the long-term need and preservation of the data.

- Section 7 describes the strategy for data sharing.
- Section 8 refers to the resources needed for the SYCLOPS data collection and management process.
- Appendix I includes a template used to collect information from all partners. We developed this form using Digital Curation Center's DMPOnline framework as a guideline[1].

[1] https://dmponline.dcc.ac.uk/

2 Data Collection

WHAT DATA WILL YOU COLLECT OR CREATE?

The SYCLOPS project will create/use three types of data:

- Open access datasets to evaluate use cases.
- Benchmarking and profiling data to measure the performance of the use case application.
- Infrastructure utilization data to measure the efficiency of the applications.

SYCLOPS has selected 3 use cases: (i) providing accelerated AI for the autonomous systems use case, (ii) scalable analysis of Petabytes of data in high-energy physics use case, and (iii) accurate analysis of heterogeneous genomic datasets in precision oncology use case. For the precision oncology use case, publicly-available, open-access genomic datasets from GIAB[2], TCGA[3], etcetera, have already been widely used for benchmarking genomic data analysis pipelines[4]. We will rely on such datasets. Similarly, research access datasets will be used for autonomous use case as well. We would also be creating a set of micro-benchmarks and synthetic datasets aiming at exploring the design space.

We also intend to collect benchmarking and profiling data for the following two use cases:

- Development of SYCL-ROOT and proving its usefulness in the context of SYCLOPS. Examples include the amount of CPU time spent by specific parts of the application, the number of SYCL tasks deployed to specific devices, the number of function calls, the amount of data transferred between CPU, GPU and RISC-V etc.

- Performance modelling in SYCL-GAL variant-calling developments, we will aim at collecting the profiling and benchmarking results on SYCLOPS-targeted computing architectures and systems.

We will also collect infrastructure utilization like energy efficiency and CPU/GPU utilization to show that SYCLOPS achieve its objectives based on predefined KPIs. The successful achievement of such objectives also relies on the production of documentation, proof-of-concept systems, performance measurements and scientific publications, all of which will be made publicly available.

HOW WILL THE DATA BE COLLECTED OR CREATED?

The collection of benchmarking data will be conducted by relying on micro-benchmarking and profiling of targeted computing architectures, while the evaluation will be based on publicly available and specifically created synthetic datasets.

Input data/datasets for the SYCLOPS use case applications:

- will be downloaded from public FTP servers or GitHub page where they are hosted.

[2] https://www.nist.gov/programs-projects/genome-bottle

[3] https://www.cancer.gov/ccg/research/genome-sequencing/tcga

[4] Intel, Deploying GATK Best Practices, https://www.intel.com/content/dam/www/public/us/en/documents/white-papers/deploying-gatk-best-practices-paper.pdf

- will use the publicly-available data from the CERN experiments (e.g. LHC experiments like ATLAS)
- will create synthetic datasets to evaluate libraries.

Various benchmarking and profiling tools will be used to obtain performance data. We will also deploy various performance & energy consumption tools to collect the infrastructure statistics & metrics.

3 Documentation and Metadata

WHAT DOCUMENTATION AND METADATA WILL ACCOMPANY THE DATA?

Data will be presented in a form understandable for a human (e.g. a spreadsheet, a graph etc.) and will be accompanied with explanations of any IDs, columns, symbols etc. used there. Code will be versioned along standard versioning conventions. Documentation will be comprised of specification descriptions, code examples and testbeds, explanations of metrics used, produced using standardised document types. Additionally, to facilitate reproduction of the results, raw benchmarking data will also be included with the reproduction instructions.

We will work with our use case partners in referencing metadata in publications. When possible, we will rely on previously established guidelines for metadata documentation particularly for use cases. For example, for the genomics use case, metadata information about publicly available datasets, like the reference assembly used, access identifiers, etecetera, and ways to catalog them are well established and used by several other benchmarking studies. We will use and report the same metadata in our study as well. Where appropriate, we will make DOI's available, pointing to data stored on repositories like Zenodo.

4 Ethics and Legal Compliance

How will you manage any ethical issues?

No data from live subjects will be used, produced, or stored. While we cannot predict how the eventually published data will be used, we will not contribute on purpose towards any ethical problem known to us.

How will you manage copyright and Intellectual Property Rights (IPR) issues?

We will work with the other consortium members and the coordinator in drafting and framing the IP ownership rules as a part of the IPR management plan.

5 Storage and Backup

How will the data be stored and backed up during the research?

All data storage and backups will be maintained either on CERN own servers, GitHub, the EURECOM Cloud (or a similar central resource for SYCLOPS), EMDC server, INSEC-ID infrastructure or servers of institutions collaborating with CERN like the Barcelona Supercomputing Center.

Scripts used for benchmarking and experimental results will be published in public repositories following regulations specified in the project management plan document.

How will you manage access and security?

All results and code for SYCL-ROOT, SYCL-DNN will be made open access after discussion of IP and exploitation possibilities when relevant. Access to any other data is based on dedicated restricted access server per partner.

6 Selection and Preservation

WHICH DATA ARE OF LONG-TERM VALUE AND SHOULD BE RETAINED, SHARED, AND/OR PRESERVED?

The following data will be of long-term value :

- Genomic data used for benchmarking, which is already preserved in public repositories.
- The data from the CERN experiments, which are already retained since their purpose is much wider than SYCLOPS.
- The benchmarking and profiling data used for proving the usefulness of libraries (SYCL-GAL, SYCL-ROOT) is also of long-term value and will be preserved.
- Scripts and data for reproducing experiments will be of long-term value.
- Source code for libraries like SYCL-ROOT, SYCL-DNN and SYCL-GAL which will be open sourced.

WHAT IS THE LONG-TERM PRESERVATION PLAN FOR THE DATASET?

The dataset will be stored inside relevant SYCLOPS resources (e.g. EURECOM Cloud) or secure CERN servers (e.g. CERN GitLab) or heiArchive service.

7 Data Sharing

HOW WILL YOU SHARE THE DATA?

Genomic data used for benchmarking is already shared in public repositories. Apart from the data which is already public, data will be shared via GitHub, the SYCLOPS mailing list and the SYCLOPS repository on EURECOM Cloud. Data required for reproducing experiments will be shared on Zenodo. EURECOM's gitlab or public github will be used to archive scripts and code to ensure reproducibility of experimentation. Further, data are reproducible from EMDC using monitoring tool and it will be stored in tool database in EMDC.

Public data is also being made available at the project's website and repositories, as appropriate. Additionally, users will also be made aware of the data through dissemination activities, invited talks, and social networks, among others.

ARE ANY RESTRICTIONS ON DATA SHARING REQUIRED?

No restrictions are required.

8 Responsibilities and Resources

WHO WILL BE RESPONSIBLE FOR DATA MANAGEMENT?

Each partner will use available resources and be responsible to prepare and manage their own data.

WHAT RESOURCES WILL YOU REQUIRE TO DELIVER YOUR PLAN?

No special resources are required. We only require support to create websites/repositories.

Appendix - I

Data Collection
What data will you collect or create?
How will the data be collected or created?

Documentation and Metadata
What documentation and metadata will accompany the data?

Ethics and Legal Compliance
How will you manage any ethical issues?
How will you manage copyright and Intellectual Property Rights (IPR) issues?

Storage and Backup
How will the data be stored and backed up during the research?
How will you manage access and security?

Selection and Preservation
Which data are of long-term value and should be retained, shared, and/or preserved?
What is the long-term preservation plan for the dataset?

Data Sharing
How will you share the data?
Are any restrictions on data sharing required?

Responsibilities and Resources
Who will be responsible for data management?
What resources will you require to deliver your plan?