SYCLOPS

Deliverable 1.3 – 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**

D1.3 – Data Management Plan

Executive Summary:	D1.3 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:	31-03-2024	**Actual Submission Date:**	02-04-2024

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	01/03/2024 – 1st draft of the deliverable	All partners
0.2	29/03/2024 – 2nd draft of the deliverable	CPLAY
1.0	02/04/2024 – 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

Statement of Originality

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Table of Contents

Executive Summary

The document provides the M15 update of the Data Management Plan (DMP) of the SYCLOPS project. The deliverable recaps the 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. At M15, we have assessed the initial DMP and re-evaluated to discover if it has been affected by future results of the work performed in all technical Work Packages. As the project evolves, we will further evolve this DMP during the project lifetime as a living document.

1 Introduction

Previously, we submitted D1.2, the initial Data Management Plan (DMP) which established a framework for the data management policy of the SYCLOPS project. The document provides the M15 update of this DMP concerning the data processed, generated, and preserved during the first 15 months of the project. In addition, any topics of discussion regarding data usage, ethics, and security are also recapitulated and discussed in this deliverable. This deliverable will be continually assessed during the project and updated accordingly.

1.1 Background

This deliverable D1.3 - 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 M15 revision of the DMP.

1.2 Purpose and Scope

The DMP 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

In order to draft this deliverable, similar to what we did for D1.2, we circulated a template using Digital Curation Center's DMPOnline framework as a guideline[1] to gather partner input. Hence, we have structured this document to mirror the initial DMP delivered in D1.2. More specifically, we describe updates based on project progress so far as follows:

- Section 1 (this section) provides the DMP'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.

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

- 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.

2 Data Collection

WHAT DATA WILL YOU COLLECT OR CREATE?

In D1.2, we identified that 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.

At M15, we confirm that these three types are still the only sources of data generation.

HOW WILL THE DATA BE COLLECTED OR CREATED?

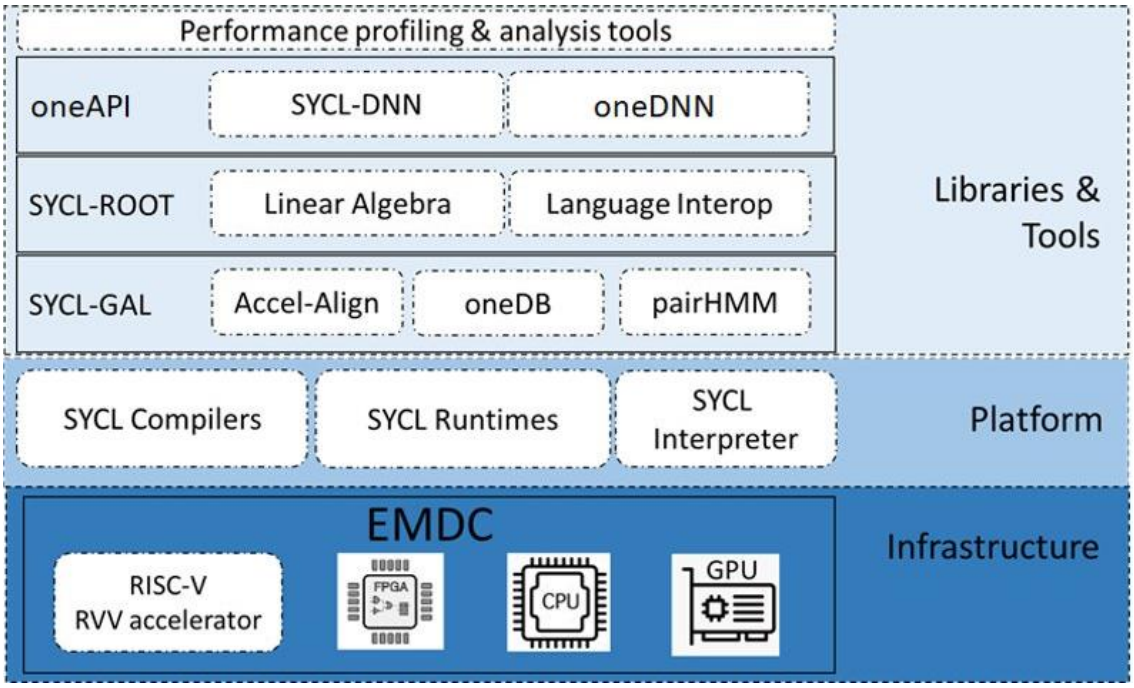

Fig 1. SYCLOPS architecture

Figure 1 above shows the core SYCLOPS hardware—software stack consists of three layers: (i) infrastructure layer, (ii) platform layer, and (iii) application libraries and tools layer. We have been developing several components in each layer till M15. Thus, here, we provide an overview of which type of data mentioned earlier (open access datasets, benchmarking, or infrastructure data) will be generated at each layer, and what has already been generated.

Infrastructure layer (RVV): The SYCLOPS infrastructure layer is the bottom-most layer of the stack and provides heterogeneous hardware with a wide range of accelerators from several vendors. A key accelerator in this layer will be the RISC-V accelerator designed by CSIP. At M15, a key update with respect to data generation is that a RISC-V emulator with support for various RISC-V extensions, named Spike, has been developed by CSIP. The emulator is used as a reference for various testing systems and is being used to generate benchmarking ad profiling data at the processor level.

In order to demonstrate (i) an end-to-end integration of open standards, and (ii) the cross-architecture, cross-vendor performance portability of SYCLOPS, our partner HIRO will package the RVV accelerator with CPU, GPU, and FPGA from several other leading processor manufactures and build modular, energy-efficient edge microdatacenters (EMDC). The EMDC will be used as a research and development testbed. At M15, we are finalizing the ordering and delivery of v1.0 of the EMDC. Once the EMDC is operational, we expect to use performance profiling tools developed in SYCLOPS to gather infrastructure utilization like energy efficiency and CPU/GPU utilization to show that SYCLOPS achieve its objectives based on predefined KPIs. We will describe the type of data in detail below.

Platform layer: The platform layer provides the software required to compile, execute, and interpret SYCL applications over processors in the infrastructure layer. SYCLOPS contains oneAPI DPC++ compiler from CPLAY, and hipSYCL, an open-source SYCL compiler toolchain from UHEI. At M15, DPC++ and hipSYCL have been substantially extended to support several new functionalities. In terms of SYCL interpreters, SYCLOPS will contain Cling from CERN. As mentioned, Cling is a state-of-the-art C++ interpreter that is being used as an interactive code development environment for exploratory analysis. At M15, Cling is being actively extended to natively support SYCL and enable Jupyter notebook-based, accelerated ad-hoc analytics. As all these components are core systems software in SYCLOPS, they do not directly contribute to data generation by themselves.

Application libraries and use cases: The libraries layer in SYCLOPS enables API-based programming by providing pre-designed, tuned libraries for our use cases. 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. At M15, we have made substantial progress with respect to each use case and its library development.

1. Autonomous systems use case

We have developed the first version of the Pointnet model in the SYCL-based portDNN library. A trained PointNet model will be used to evaluate the Autonomous systems use case. We will be training the model with ModelNet40[2] dataset which is a publicly available dataset.

2. High-energy physics use case

We have developed GenVectorX, a C++ package that provides classes and functionalities to represent and manipulate particle events using both CUDA and the SYCL programming model. We have generated simulated data to drive Lorentz vector computations. GenVectorX has been evaluated locally at CERN on AMD and Intel CPUs and NVIDIA GPUs. As the library provides acceleration of key compute kernels, there is no specific data that is of relevance to the DMP. We will use it in integrated use case testing in the future.

3. Precision oncology use case

At M15, we have developed a single-threaded version of all stages of the GATK post-alignment preparation pipeline as a first step towards developing a SYCL-based parallel version. We have used the publicly-available sample data provided by NVIDIA Clara Parabricks[3] to test the single-threaded code and verify its conformance with GATK. We have also analysed the Pairwise Hidden Markov Model (pairHMM) implementation of Intel

[2] https://modelnet.cs.princeton.edu/
[3] https://docs.nvidia.com/clara/parabricks/4.0.0/tutorials/gettingthesampledata.html

Genomics Kernel library and developed several optimizations to it that have been tested using publicly available test data provided by Intel as a part of the codebase[4]. In the upcoming months, we plan to use the publicly-available, Genome-In-a-Bottle[5] human whole-exome sequencing dataset NA12878 to do end-to-end testing at the use case level.

In the upcoming months, we will be working on a preliminary, interim end-to-end evaluation of all use cases using our EMDC v1.0 with the main goal of proving compatibility and integration. At this stage, we will be using profilers where relevant to gather data about various aspects, including 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 accelerator etc. We will also be creating a set of micro-benchmarks and synthetic datasets aiming at exploring the design space of each use case. All data that will be used for each use case will be publicly available.

Performance profiling tools and analysis: The libraries layer also includes tools that we are developing for performance profiling and analysis. At M15, we have developed AdaptivePerf. an open-source, comprehensive, low-overhead code profiler that does not require any code instrumentation. AdaptivePerf can generate several types of profiling data including the following:

- Profiling exact runtime length of each thread/process spawned by a profiled program (including the main thread) to be displayed on a graphical timeline.
- Sampling on-CPU times of each process/thread and displaying them in form of non-time-ordered and time-ordered flame graphs along with exact off-CPU times.
- Sampling off-CPU times of each process/thread to be displayed on a timeline.
- Obtaining stack traces of a function spawning each thread/process.
- Sampling any user-provided perf event and displaying its occurrences in form of non-time-ordered and time-ordered flame graphs.

We are actively improving AdaptivePerf and plan to use it to profile the SYCLOPS libraries in the upcoming months. Finally, on the performance modelling front, we have developed a new methodology based on Cache-Aware Roofline Modelling for the scaling of hardware resources, optimizing the architecture for a given application. We tested our methodology using SpMV with publicly-available SuiteSparse matrices [6] as the target application. Thus, with respect to DMP, we rely only on public data as input and we don't generate any downstream ``usable'' data that must be protected as output.

[4] https://github.com/Intel-HLS/GKL/tree/master/src/test/resources
[5] https://www.nist.gov/programs-projects/genome-bottle
[6] S. Kolodziej *et al.*, "The SuiteSparse Matrix Collection Website Interface," *JOSS*, vol. 4, no. 35, p. 1244, Mar. 2019, doi: 10.21105/joss.01244

3 Documentation and Metadata

WHAT DOCUMENTATION AND METADATA WILL ACCOMPANY THE DATA?

As described in Section 2, the types of data relevant to SYCLOPS are:

- 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

Regarding open access data, our use case partners already rely on rely on previously established guidelines for metadata documentation. For all our cases, metadata information about publicly available datasets, for instance like the reference assembly used, access identifiers, etcetera, in genomics, and ways to catalogue them are well established and used by several benchmarking studies. For example, we reported earlier in this document an example ID from the GIAB dataset that we plan to use. In all reports derived from such open-access data, we will use and report such all relevant metadata in all publications with relevant DOIs. For the autonomous systems use case, an important metadata is the trained model itself. This model will be hosted in an accessible location where they can be accessed by anyone who would like to reproduce the analysis results.

Regarding benchmarking data, we will include detailed instructions in each code base together with automated scripts to facilitate reproduction of our benchmarking results by others.

Infrastructure utilization data is primarily used to verify the effectiveness of our stack in exploiting heterogeneous hardware. This data, in addition to benchmarking data and other results, will be presented to scientific audience in publications and also in a simplified form understandable by the general public in blogs, videos, and other dissemination sources.

4 Ethics and Legal Compliance

How will you manage any ethical issues?

We confirm at M15 that no data from live subjects will be used, produced, or stored.

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

We have worked with all consortium members in drafting and framing IP ownership rules as a part of the IPR management plan. A first version of the IPR plan was released in deliverable “D6.2. SYCLOPS IPR Management, Business Models, and Business Plan (M12)”.

5 Storage and Backup

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

As described in D1.2, each partner site has storage facility for storing data relevant to their research either on site, or at partner institutions. There is no change of plan with respect to data storage and backup.

How will you manage access and security?

D1.2 described that in general, all libraries and tools developed in SYCLOPS at the platform and libraries layers will be made open access after discussion of IP and exploitation possibilities when relevant. Access to any other source, binaries, and/or data (like CSIP solutions from infrastructure layer) is based on dedicated restricted access server per partner. There is no change in this at M15.

6 Selection and Preservation

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

We identified the following data will be of long-term value in D1.2:

- 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.

At M15, we have no changes to this list.

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

We have deployed at EURECOM Yeti, a 20PB archival storage service[7]. All data that is deemed as valuable for long-term storage will be stored here. We will also use CERN servers (e.g. CERN GitLab) or heiArchive service from UHEI if appropriate as explained in D1.2.

[7] https://yeti4d.ds.eurecom.fr/

7 Data Sharing

How will you share the data?

There is no change in the plan for data sharing from D1.2. As described earlier, use cases already use publicly-available datasets. As all new libraries and tools developed in SYCLOPS are open source in nature, we have created a central SYCLOPS GitHub repository that will pool together all relevant code and make it available under one roof. In addition, code bases will also exist in partner-specific git repositories to ensure their continued support after the project. Apart from this, data will be shared via SYCLOPS mailing list and the Yeti SYCLOPS repository. 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. No changes at M15.

WHAT RESOURCES WILL YOU REQUIRE TO DELIVER YOUR PLAN?

No special resources are required as of M15. 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?