D1.2 Data Management Plan (DMP)
Version 1.0

Document Information

Contract Number	824080
Project Website	www.pop-coe.eu
Contractual Deadline	M6, May 2019
Dissemination Level	PU
Nature	R
Author	BSC
Contributor(s)	BSC
Reviewer	USTUTT-HLRS
Keywords	Data management, analysis reports, website, co-design activities, FAIR data, GDPR

Notices:
The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No "824080".

Change Log

Version	Author	Description of Change
V0.5	BSC	Initial Draft
V1.0	BSC	Changes after POP internal review by USTUTT-HLRS.

Table of Contents

Acronyms and Abbreviations

- BSC – Barcelona Supercomputing Center
- CoE – Center of Excellence
- D – Deliverable
- DMP – Data Management Plan
- EC – European Commission
- GA – General Assembly
- GDPR - General Data Protection Regulation
- GitLab – Open source application for the entire software development lifecycle
- GPL - General Public License
- HPC – High Performance Computing
- LGPL - Lesser General Public License
- NDA – Non Disclosure Agreement
- POP – Performance Optimization and Productivity
- TRAC – Wiki and issue tracking software
- USTUTT-HLRS – University of Stuttgart - High Performance Computing Center Stuttgart
- URL – Universal Resource Locator
- WP – Work Package

Executive Summary

This deliverable analyses the main elements of the data management policy with regard to all the datasets collected, processed and/or generated along the lifetime of the project "Performance Optimization and Productivity 2" (POP2) and even after the project is completed.
It describes what data will be collected or generated, according to which methodology and standards, whether and how this data will be shared and/or made openly accessible, and finally how it will be curated and preserved.
According to FAIR principles for H2020 projects[i], this document outlines the data management life-cycle for all datasets in order to make research data findable, accessible, interoperable, and reusable (FAIR).
The Data Management Plan is a living document. POP management team is responsible for updating this document, collecting feedback from the GA members and the BSC data management experts. The document will be updated as soon as any relevant modification is implemented.

1. Background

POP2 is operating a Centre of Excellence (CoE) in Performance Optimisation and Productivity of Parallel Applications extending the activities of the previous POP project (www.pop-coe.eu).
The mission of POP2 is to promote best practices in performance analysis and optimization, helping developers and users of parallel applications to understand the performance of their applications and identifying ways to improve them, with the final goal of improving their productivity and competitiveness.
POP2 offers a portfolio of services designed to help users optimise parallel software and understand performance issues. The primary customers are code developers and owners. However, POP2 services are also available to code users, research infrastructure and service centres.
The main role of POP2 is to offer and implement these free services to EU institutions helping a wide user community in the performance optimisation of their codes. The research of POP2 is limited to WP7 (co-design activities) and WP8 (tools and methodology).

2. Data classification and analysis

The data managed by POP2 can be classified into these types:

1. Data from POP2 customers (WP2 and WP3)
2. Data generated for marketing and dissemination activities (WP2 and WP4
3. Data from POP2 assessments (WP5)
4. Data from POP2 Proof-of-Concept (WP6)
5. Data generated by WP7 with respect to the co-design activities
6. Data related with WP8 where the tools and methodology are improved

The treatment of types 1, 2 as well as the personal data collected by 3 and 4 is stored in the TRAC system. This has been described in deliverable D9.1[ii] which reports also on the ethical issues. Therefore, they will not be covered again in this document.
We include a subsection for each of the other types, despite the main focus is given to the last type as it is the one related with research activities so it would be a target for the FAIR.

2.1 Data from POP assessments

The input for a POP assessment is the performance data that may be collected by the user that request the service or by the POP partner. In this second case, POP would need access to at least the binary and the input files. As these files belong to the user of the service and the access to them would be temporal, we can consider that the first data managed by the project is the performance data collect.
The performance data is stored in the machines used by the POP partner to do the analysis. The only requirement is to keep the data while it is foreseen to have further activities with the same user. The performance data can be shared between partners unless the study required an NDA. In any case, the data will not be distributed outside the consortium without previous consent from the user.
As output of the analysis, we produce a report (slides or document) with the results and observations. These documents are stored in the TRAC and wiki. If the user grants us permission, they are also published in the website.

2.2 Data from POP Proof-of-Concept

The POP Proof-of-Concept is a second step service where the code (or a mock-up) is modified to demonstrate the potential benefits of the suggested improvements. The modifications can be done by the POP partner or directly by the user with the support of POP. In both cases, these modifications are specific for the code and there is no further treatment. In POP2 we have a new work package that would implement Co-design activities based on the experience gathered in the PoC services.
The performance data required during the analysis and evaluation is stored in the machines used by the POP partner to elaborate the PoC. As previously mentioned, the only requirement is to keep the data while it is foreseen to develop further activities with the same user. The performance data can be shared between partners unless the study required an NDA. In any case, the data will not be distributed outside the consortium without previous user consent.
As output of the PoC, we produce a report (document) with the results and observations. These documents are stored on the TRAC and wiki. If the user grants us permission, they are also published in the website.

2.3 Improvements in the tools and methodology

In POP2, there is a work package with the specific goal of improving the methodology used for the analysis as well as to enhance the tools developed by some of the partners to be more effective for using them during the analysis.
All the tools developed by the partner are open source and have well stablished channels for distribution. The target would be to include on the official releases the improvement implemented during the project as soon as they are successfully validated. We consider there is no need to create a new site to store these releases.

2.4 Co-design repository

The co-design repository is a centralized database for source code and metrics associated with it. Its main goal is to provide the HPC community with a set of kernels which potentially could be used for training, dissemination and system/software design activities.
The kernels will be managed through the GitLab Community Edition, which is open source software. The GitLab repository will be managed by the BSC personnel to control the creation of individual source code repositories and will nominate a maintainer who will be responsible for assigning additional permissions to the rest of developers (per-user access control).
Maintainers are free to follow their versioning conventions, but all public releases must be tagged with their version. The metrics to include in the repository will refer to this version. Source code repository maintainers will use branches as needed or required.
In D7.1[iii] the co-design repository structure is described and defines a set of guidelines in order to document the source code. The management of the collections will be done through changes in this repository. Kernel codes and the meta-data associated with them will live in the corresponding repository. In the current design, kernel's meta-data consist of the following collections: programs, versions, experiments and results.
Private access during analysis development guarantees access only for POP2 team members. Once the reporting accomplishes the minimum requirements, public access will be provided on the project website.
Initially, the GitLab repositories will be backed up daily. If the backup size becomes large, we will switch to a daily incremental backup policy.
POP2 does not pass personal data on to third parties and would obtain additional explicit consent to any secondary use of the data as described in D1.1[iv] in compliance with the GDPR[v]. Before any code release, partners should have received the customer´s approval.
The final goal is to have a permanent repository to be queried beyond the project duration.

3. FAIR Data

Relevant results of the project with respect to the POP methodology will be published on the project website, the co-design repository, other institutional repositories, Zenodo (https://zenodo.org/) or in open access scientific publications.

Main documents (performance assessment and proof-of-concept reports) are described in Section 2, but only those with previous customer consents will be finally published on the website. These documents follow an internal naming convention that still remains when published in the website:

- Performance assessment: POP2_AR_nnn (where “nnn” is a unique identifier among all assessment reports).
- Proof-of-concept reports: POP2_PoCR_nnn (where “nnn” is a unique identifier among all assessment reports).

A subgroup of these documents will be also published under the website section “success stories” (https://pop-coe.eu/target-customers/success-stories) where the POP service activity will be detailed. Success stories will also link to the original corresponding report.

The co-design repository may also publish results derived from a performance assessment or proof-of-concept public reports. In these cases, the co-design website will also link the original document by means of its URL.

All the kernels included in the co-design repository will contain a license file describing the terms of use. The project will encourage that all of them are based on GPL or LGPL licenses in order to avoid any kind of restriction on terms of use and distribution.

Another option are the resulting papers where lessons learned from multiple similar performance assessments will be summarized and documented for outside readers.

[i] European Commission. Guidelines on FAIR Data Management in Horizon 2020, Version 3.0, 26 July 2016.
[ii] Deliverable D9.1 POPD – Requirement No. 1 – Ethics issues, submitted on February 28 2019
[iii] Deliverable D7.1 Co-design repository structure, submitted on May 31 2019
[iv] Deliverable D1.1 Project Management and Quality Guidelines, submitted on February 28 2019
[v] General Data Protection Regulation (EU) 2016/679 ("GDPR")