HPC/Exascale Centre of Excellence in Personalised Medicine

D3.2 Data Management Plan Version 1.0

Document Information

Contract Number	951773
Project Website	http://www.permedcoe.eu/
Contractual Deadline	M6, March 2021
Dissemination Level	PU
Nature	R
Author(s)	Sarah Peter (UNILU), Wei GU (UNILU), Christophe Trefois (UNILU)
Contributor(s)	Laurence Calzone (IC)
Reviewer(s)	Arnau Montagud (BSC) Jesse Harrison, Sampo Sillanpää, Henrik Nortamo (CSC)
Keywords	DMP, data management plan, HPC

Notice: *The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No "951773".*

Change Log

Version	Author	Date	Description of Change
V0.1	Wei Gu	2021-01-28	Initial Draft
V0.2	Sarah Peter	2021-02-09	Added details about use cases
V0.3	Sarah Peter	2021-03-01	Clean version ready for internal comments
V0.4	Sarah Peter, Laurence Calzone	2021-03-09	Review by Laurence
V0.5	Sarah Peter, Arnau Montagud, Jesse Harrison, Sampo Sillanpää, Henrik Nortamo	2021-03-18	Review by Arnau, Jesse, Sampo and Hendrik
V1.0	Sarah Peter	2021-03-19	Clean-up
V1.0	Alba Jené	2021-03-30	Final edits before submission to the EC

Table of contents

Executive Summary

This document reports the Data Management Plan (DMP) of datasets used in the PerMedCoE use cases. It is developed on the basis of a survey conducted among the consortium partners regarding the datasets that will be used in the use cases, how the data will be stored, accessed, used, and updated throughout the length of the project. The use case coordinating team will review, coordinate, and resolve issues related to the DMP and update it accordingly. The DMP will be a living document and updated annually. The current version of DMP summarizes the volume, types, and source of different datasets. It also reports on, wherever applicable, data standards, data security, access control, FAIRification - making data Findable, Accessible, Interoperable, Reusable (FAIR), and long-term sustainability plans.

1. Introduction

This deliverable is to report the Data Management Plan (DMP) of datasets used in the PerMedCoE use cases. The DMP will document the facts of the datasets, how they will be collected, stored, curated/standardised, processed/analysed and how the access and sharing will be governed, how the data will be reserved, protected and destructed (if needed). This DMP development is linked to the task 3.2 in the PerMedCoE project.

Task 3.2. Use cases to be performed in pre-exascale architecture [M1 – M36] (Lead: IC – Participants: BSC, IRB, CSC, UNILU, UKHD, IBM, KTH)

Task leader team: IC

Description of work:

This task will target the application of the optimised frameworks and developed workflows to real PerMed use cases, running in pre-exascale environments. Five use case scenarios are presented to showcase the impact of the PerMedCoE's developments. To ensure privacy requirements compliance, cross-WP teams will be established per use case including members of WP1, WP2, T3.1 and T5.1, in which privacy experts will support the coverage of responsibilities under the GDPR in the different stages of technical developments. The outputs from such cross-WP teams work will be reported as the Data Management Plan (D3.2), which will be kept updated as the project evolves. The use case reports will also include ethics clearance documentation and Mutual Transfer Agreements (MTAs) as required for each particular scenario (reported in D3.8).

The DMP is a document which adequately reflects the current state of the datasets used in the project, yet will be continuously updated throughout the course of the project, as e.g., data formats or storage requirements may change during the acquisition of data.

2. Description of Activities

This deliverable is a result of the following steps:

1) Conducted a survey of information on datasets from all use cases.
2) Used the use case coordination meeting to sync information and resolve questions/issues as well as updating future versions of the DMP.

3. Results

3.1 Description of the data

3.1.1 Type of study

Use Case 1 will propose cancer treatments for individual patients using associated clinical information employing omics data and personalised cell-level models.

Use Case 2 will find effective drug combinations for cancer by using drug-response experiments and publicly available databases, personalised cell-level models and BioExcel's GROMACS simulations.

Use Case 3 will focus on simulations of individual disease characteristics, such as the degree of severity, in support of the rare diseases' diagnoses where only a small number of cases are available.

Use Case 4 will perform simulations of millions of individual cells, each one of them with their own omics data and individual models of metabolism, regulation and signalling. Cells will interact with each other (agent-based models) and with the environment, modelled as fluids, from where cells get nutrients and input information, in gradients related to their relative position to the other cells.

Use Case 5 consists of modelling the cell types and intracellular pathways that have been identified as playing a role in COVID-19 disease.

The detailed study (analysis) plan of the use cases is reported in Deliverable D3.1.

3.1.2 Type of data

Types of data generated and collected in the use cases include:

- Single-cell RNA-seq (UC5)
- (Bulk) RNA-seq (UC1, UC2, UC5)
- Single cell transcriptome and epitope profiling of mouse salivary glands using Drop-seq and CITE-seq (UC4)
- RNA arrays (UC2)
- Somatic mutations, copy number alterations, germline variants, DNA methylation and expression (UC1)
- Reverse Phase Protein Array (RPPA) data (UC2)
- Cell viability data (UC2)

This list will be updated with new prospective data types from the different use cases and will be provided in future releases of the DMP.

3.1.3 Format and scale of the data

The format and scale of datasets in this project are summarized in Table 1. At this stage of the DMP, the size of each dataset is roughly estimated, and details of the

format and size will be updated during each revision (new version) of this DMP when more information is available.

Dataset	Format	Size
GSE148729 (UC 5)	Tabular	300 MB
GSE124425 (UC4)	SAM, tabular, text	160 GB
GDSC gene expression (UC2)	Tabular, text	4 GB
CLL-ICGC (UC1)	MAF	9 GB
NTNU public dataset (UC2)	Tabular	8 MB
NTNU RNA-seq private dataset (UC2)	Tabular	< 1 GB
NTNU RPPA private dataset (UC2)	Tabular	< 1 GB

Table 1: Format and size of datasets

Thus, the expected storage volume is in a non-critical range.

3.2 Data collection / generation

3.2.1 Methodologies for data collection / generation

We will collect the datasets from public resources and various consortium partners using a secure communication channel (see D5.6 for details) into Data and Computing Platforms hosted at IC, BSC, UKHD and UNILU. In addition, we will collect the details of data collection systems and data formats for each data type from each data collection site via an online Electronic Data Capturing system (REDCap)[1] and will provide updates regularly in future versions of DMP.

Besides the data, associated scientific metadata as well as data-protection-related metadata to fulfil GDPR[2] requirements (depending on each data type) will also be collected at the data collection site.

3.2.2 Data quality and standards

Since data are not generated within this project, the data quality and standards will follow the *status quo* of those from the (public) resources from which the data are obtained, e.g., GEO and ArrayExpress.

[1] https://www.project-redcap.org/
[2] https://gdpr-info.eu/

Data standards for each dataset used in the use cases will be evaluated during the project to reflect the updates from the use case coordinating group.

3.3 Data management, documentation, curation and FAIRification

3.3.1 Uploading, managing, storing and curating data

Each partner providing data for use cases will act as a data controller. They will provide decisions from their local authorities for data storage before the next update of the DMP. If data need to be transferred to another site for processing, the data provider will upload their data using a secured channel (see D5.6 for details). A detailed list of requirements will be reported in D3.9.

Use case	**Data source**	**Data storage sites**
UC5	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE148729	IC, BSC, UKHD
UC4	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE124425	BSC, IC, UKHD, UNILU
UC2	https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-3610/	IC, UKHD, BSC, UNILU
UC1	https://dcc.icgc.org/releases/current/Projects/CLLE-ES	BSC
UC2	https://figshare.com/articles/dataset/A_high-throughput_drug_combination_screen_of_targeted_small_molecule_inhibitors_in_cancer_cell_lines_-_supplementary_table/9810719	IC, UKHD, BSC, UNILU
UC2	Asmund Flobak, NTNU	IC, UKHD, BSC, UNILU

Table 2: Summary of data sources and data storage sites

The resulting data will be made available for use case members, once adequate ELSI (ethical, legal and social implications) processes are in place. Within the consortium, all partners will follow their own ethical protocols and informed consent. As a best practice, we will encourage partners to re-pseudonymise any sensitive data before they are stored and processed.

The FAIRness of all the research data will be improved by applying the principles of **findable, accessible, interoperable and reusable (FAIR)** and using the tools/cookbooks developed by the IMI-FAIRplus project[3] as broadly as possible.

These principles are implemented in PerMedCoE use cases by ensuring that data follow these rules:

- Datasets need to be assigned a unique identifier within the project. The data management team will ensure that the identifier is globally unique. We propose using a standard such as UC#-#. For instance, UC2-3 would be the third chronologically acquired dataset of use case 2.
- Accompanying data such as the study protocol, experimental parameters etc. are provided. This will make it possible for consortium members to fully grasp the experimental setup and data content.

3.3.2 Metadata standards and data documentation

As the datasets used in the use cases are either from existing public sources or only meant for the testing of the development of software and workflows, the consortium does not aim at publishing the datasets (as well as its associated metadata). Therefore, all metadata and data documentation will be kept in the same format as they were obtained from the public database or data provider.

3.3.3 Data preservation strategy

The data controllers of each dataset will, if resources permit, store a local copy at their sites. Data/files hosted at the processing site will be backed up to minimize the risk of data loss. There will be no long-term storage of data within PerMedCoE beyond the duration of the project.

3.4 Data security and confidentiality of potentially disclosive information

3.4.1 Formal information/data security standards

Security measures that will be implemented in the use cases will be reported in D5.6.

Personal data for administrative / communication purposes that is being processed through BSC is only collected to perform the tasks the data controller (BSC) is obliged to. Therefore, the data processed is exclusively useful for the employees involved in the project activities. Personal data is only processed for the purposes described in the Grant Agreement and while maintaining the appropriate technical and organizational measures.

Data collected within event registration or the use of the websites related to PerMedCoE is inaccessible for other partners and third parties. With regard to the use of the PerMedCoE website and the publication of audio and visual recordings on the

[3] https://fairplus.github.io/the-fair-cookbook/content/home.html

Internet, BSC is aware that data transmitted via the internet may be subject to security breaches. Complete protection from third-party access is not possible.

Appropriate technical and organizational measures have been identified and implemented together with the Data Protection Officer (See Annex 1, DPOs). BSC is located in Spain. The data is stored in Spain as well.

3.4.2 Data Protection Impact Assessment

The data protection impact assessment (DPIA)[4] will be reported in D3.9.

3.4.3 Ethics

All ethics and legal assessment of use cases is reported in D3.8.

3.5 Data sharing and access

3.5.1 Data sharing

Data will only be shared within the consortium, as detailed in 3.3.1.

3.5.2 Findability of the research data by potential users during/beyond the project

Data will be made available to the use case participants in accordance with the Consortium Agreement (CA) and GDPR.

Data will not be (re-)distributed outside the consortium.

3.5.3 Governance of access

The Data Use and Access Committees (DUACs) of each dataset are the only bodies that will be able to grant or revoke access. The hosting institutes of the data sharing platforms (servers) are responsible for the technical implementation of access control based on the decision of the data owner or DUACs.

Internal access: The data as well as data/results generated within a given use case will be open to all the use case participants in accordance with the use case work plan (D3.1 and further updates). The data will be shared after concluding the appropriate data sharing agreements in compliance with Art. 26 & 28 of the GDPR.

Change/Revoke access: For institutional data access, the access rights will be managed by the DUACs of each dataset, which can change and revoke access upon request (e.g., by the Data Provider). In addition, the committees should regularly review the access rights. Internal validation should be performed periodically requesting the PI from each partner institute as well as the data controller to confirm the current user list and access rights. For individual data access, the PIs from each

[4] https://gdpr.eu/data-protection-impact-assessment-template/

partner institution are responsible to inform the sysadmin team of any changes or revoke of access of their team members at least one month in advance.

External access: There will be no access provided by the consortium to the data for external users. External users will be able to access the data from the public repositories or the data owner, according to their respective guidelines.

3.5.4 The study team's exclusive use of the data

The datasets collected in this project will only be used according to the DoW / Consortium Agreement and within the scope of the informed consent.

3.5.5 Restrictions or delays to sharing, with planned actions to limit such restrictions

The datasets should be shared with all partners once the legal procedures (informed consent, CDA, MTA, ethical approval, DSA) are in place and following the workflow defined in 3.5.1.

3.6 Data retention and destruction

In case any study participant decides to withdraw his or her informed consent or the ethical committee of the partner institute in which the data is collected decides this data should not be included in the project, the Data Use and Access Committee (DUAC) and Data Controller of the relevant dataset will send a destruction request in written form listing the detailed information for the data related to the participant(s) within 10 working days of this withdrawal. The data hosting partner will remove all copies of the data derived from this participant and data generated from related bio-samples will be removed from all hosting systems.

If the research project ends for any reason, each Data Controller will be able to request the return of its personal data and their complete cancellation from any tool used, unless a long-term archiving is agreed between the Data Controller(s) and Data Processor(s).

In case a study participant's informed consent or a study's ethical approval expired, the data destruction process as described above will be implemented. If the dataset needs to be used beyond this expiration, the committee will decide either to request an extension of the approval to the ethical committee or to completely anonymize the datasets if the informed consent or legal basis and ethical committee approves. In the second case (anonymizing datasets), the non-anonymized data will be destroyed as described in the above data destruction process.

3.7 List of Responsibilities/Institutes

The roles and responsibilities of each participating partner and their institutes will be described in an 'additional data sharing agreement'. This additional data agreement covers provisions of GDPR Articles 26 & 28 and describes the roles and responsibilities

of data providers (controllers), joint controllers and data processors towards data subjects, authorities, etc. Execution of this additional data agreement is currently ongoing.

3.8 Updates of this Document

This DMP is a living document and will be reviewed and updated yearly by the use case coordination group.

4. Conclusion

The first version of the data management plan has been set up and agreed among the use case coordination group.

5. Deviation

There are no deviations from the Description of Action.

Annex I. DPOs

The DPOs are listed as the following:

P1- BSC:
Data Protection Officer
Barcelona Supercomputing Center
Jordi Girona no. 31
08034 Barcelona (Barcelona)
e-mail: dpo@bsc.es

P2- CSC (processor):
Data Protection Officer
CSC - IT Center for Science Ltd.
FI-02101 Espoo
Finland
Telephone : +358 9 457 2001
E-mail : privacy@csc.fi

P3-UNILU:
Sandrine Munoz
Central Administration
Université du Luxembourg
Maison du Savoir, E14 1425-100
2, Avenue de l'Université
L-4365 Esch-sur-Alzette
Luxembourg
Telephone: +352 46 66 44 9813
E-mail: dpo@uni.lu

P4 - IC
Astrid Lang
26, rue d'Ulm
75005 Paris
E-mail: dpo@curie.fr

P5 – MDC
Ulrike Ohnesorge
E-mail: Datenschutz@mdc-berlin.de
www.mdc-berlin.de

P6 – UKHD
Universitätsklinikum Heidelberg
Datenschutzbeauftragte
Im Neuenheimer Feld 672
69120 Heidelberg
Telephone: +49 (0) 6221 56-7036
E-mail: datenschutz@med.uni-heidelberg.de

P7 – IRB
Data Protection Officer
C/ Baldiri Reixac, núm. 10 (Parc Científic de Barcelona)
08028 Barcelona
E-mail: dataprotection@irbbarcelona.org

Acronyms and Abbreviations

- CA – Consortium Agreement
- D – deliverable
- DMP – Data Management Plan
- DoA – Description of Action (Annex 1 of the Grant Agreement)
- DoW – Description of Work
- DPIA – Data Protection Impact Assessment
- DPO – Data Protection Officer
- DUAC – Data Use and Access Committee
- EC – European Commission
- EDC – Electronic Data Capturing
- EDPS – European Data Protection Supervisor
- EGE – European Group on Ethics
- EPF – Forum Européen des Patients
- FAIR – Findable, Accessible, Interoperable, Reusable
- GA – General Assembly / Grant Agreement
- GDPR – General Data Protection Regulation
- HIPAA – Health Insurance Portability and Accountability Act
- HPC – High Performance Computing
- HTTPS – Hypertext Transfer Protocol Secure
- IPR – Intellectual Property Right
- KPI – Key Performance Indicator
- M – Month
- MIG – Minimum Information Guidelines
- MTA – Material Transfer Agreement
- MS – Milestones
- PM – Person month / Project manager
- SCP – Secure Copy Protocol
- SOP – Standard Operation Procedure
- UC – Use Case
- UNILU – Université du Luxembourg
- WP – Work Package
- WPL – Work Package Leader