Funded by the European Union

AVITHRAPID PROJECT

101137192- AVITHRAPID

Antiviral Therapeutics for Rapid Response Against Pandemic Infectious Diseases

WP7 FAIR DATA MANAGEMENT

D7.2 Updated Data Management Plan

Lead contributor	1-FRAUNHOFER
Other contributors	All consortium partners

Data	**Description**
Due date	31-12-2024
Delivery date	30-12-2024
Delivery type	DMP
Dissemination level	PU

Document History

Version	**Date**	**Description**
V0.1	29-11-2024	First Draft – Yojana Gadiya
V1.0	23-12-2024	Final – Leonie von Berlin

Summary

The initial Data Management Plan (DMP), as outlined in Deliverable 7.1, provided foundational guidance with minimal details on folder organization and adherence to the FAIR principles (Findable, Accessible, Interoperable, and Reproducible). Recognizing the need for more comprehensive data governance, the updated version of the DMP has been significantly enhanced to address critical gaps and improve usability.

In the revised DMP, we have placed a stronger emphasis on cataloging experimental workflows and their corresponding data file formats. This systematic approach not only facilitates better organization and traceability but also ensures that data remains comprehensible and reusable across different stages of the research lifecycle.

Moreover, the enhanced DMP includes the generation of generic guidelines to streamline data handling practices. These guidelines encompass the identification and utilization of domain-specific data repositories, ensuring that datasets are stored in appropriate repositories that support long-term accessibility and interoperability. By doing so, we aim to maximize the visibility and utility of the data within the research community while aligning with best practices in open science.

The updated DMP also incorporates robust metadata standards and documentation practices to enhance data discoverability and ensure compliance with FAIR principles. These improvements reflect our commitment to fostering a research environment that prioritizes transparency, collaboration, and reproducibility. The updated DMP has been annexed with this report.

A periodic revisiting of the DMP would take place to cater towards the growing experiments and data formats across the consortium.

Antiviral Therapeutics for Rapid Response Against Pandemic Infectious Diseases

Horizon Europe
Data Management Plan

29 November 2024

History of changes

Version	Publication date	Changes
Draft v3	29 Nov 2024	First version for 6 month milestone
Draft v2	4 July 2024	Modified version based on comments from WP leads
Draft v1	20 June 2024	Initial version of DMP

Contributors

The following contributors are related to the project of this DMP:

- Yojana Gadiya
  Yojana.Gadiya@itmp.fraunhofer.de, ORCID: 0000-0002-7683-0452
  Roles: Work Package Leader
  Affiliation: **Fraunhofer Institute for Translational Medicine and Pharmacology** (ITMP)

- Björn Windshügel
  bjoern.windshuegel@itmp.fraunhofer.de
  Roles: Project Manager, Work Package Leader
  Affiliation: **Fraunhofer Institute for Translational Medicine and Pharmacology** (ITMP)

- Leonie von Berlin
  leonie.von.berlin@itmp.fraunhofer.de, ORCID: 0000-0002-7790-0395
  Roles: Data Manager, Data Steward
  Affiliation: **Fraunhofer Institute for Translational Medicine and Pharmacology** (ITMP)

Projects

We will be working on the following projects and for those are the data and work described in this DMP.

Antiviral Therapeutics for Rapid Response Against Pandemic Infectious Diseases

Acronym
AVITHRAPID

Start date
2024-01-01

End date
2028-06-30

Funding
HORIZON EUROPE Framework Programme (European Union): 101137192 (granted)

The European Consortium "Antiviral Therapeutics for Rapid Response Against Pandemic Infectious Diseases" (AVITHRAPID) aims to support the search for novel broad-spectrum antiviral compounds by advancing multiple approaches. Building on a pre-existing set of bioactive small molecules, which are at least at the validated hit level, AVITHRAPID strives for the development of pre-clinical candidates targeting several viruses. This will be achieved by combining the relevant expertise for pre-clinical drug discovery, including molecular modeling, biochemical and cell-based assays, X-ray crystallography, medicinal chemistry, biophysical binding studies, ADMETox profiling, in vitro and in vivo PK, as well as animal disease models. In addition, the consortium aims to conduct a Phase 2a clinical trial for a small molecule developed against Zika virus. Moreover, the consortium aims to identify and validate further viral targets and thereby contribute to the search for novel antiviral targets. As a consequence of the activities in AVITHRAPID, an early-stage drug discovery pipeline will be established that can be used to rapidly identify and develop novel antiviral compounds against emerging diseases.

1. Data Summary

Instrument datasets

The following instrument datasets will be acquired in the project:

- **Compound Toxicity**

  This dataset will be collected by experts in the project, with our own equipment.

  The equipment is less well described or not completely standard, so we will need to take extra care documenting the process.

  Researchers working in other fields of research could be interested in using this data. We think that other researchers can use this data as follows: The data could be used to enhance knowledge of toxicity profile of compounds.

- **Compound Activity**

  This dataset will be collected by experts in the project, at a specialized infrastructure.

  The equipment is less well described or not completely standard, so we will need to take extra care documenting the process.

  Researchers working in other fields of research could be interested in using this data. We think that other researchers can use this data as follows: The data could be used to enhance knowledge of activity profile of compounds.

- **Antiviral activity**

  This dataset will be collected by experts in the project, with our own equipment.

  The equipment is very well described and known.

  Other researchers working in the same field of research could be interested in using this data.

- **Macromolecular crystallography on protein-inhibitor complexes**

  This dataset will be collected by experts in the project, at a specialized infrastructure.

  The equipment is very well described and known.

  Other researchers working in the same field of research could be interested in using this data.

- **Liposome analysis**

This dataset will be collected by experts in the project, with our own equipment.

The equipment is less well described or not completely standard, so we will need to take extra care documenting the process.

This data are expected to be used only by us.

- **Compound Permeability**

  This dataset will be collected by experts in the project, with our own equipment.

  The equipment is less well described or not completely standard, so we will need to take extra care documenting the process.

  Other researchers working in the same field of research could be interested in using this data.

- **SARS-CoV2 model zootechnical parameters**

  This dataset will be collected by experts in the project, at a specialized infrastructure.

  The equipment is very well described and known.

  This data are expected to be used only by us.

- **Biochemical & hematology analysis**

  This dataset will be collected by experts in the project, at a specialized infrastructure.

  The equipment is very well described and known.

  Other researchers working in the same field of research could be interested in using this data.

- **Histological analysis**

  This dataset will be collected by experts in the project, at a specialized infrastructure.

  The equipment is very well described and known.

  Other researchers working in the same field of research could be interested in using this data.

Re-used datasets

We have found the following reference datasets that we have considered for re-use:

- ChemFinder

It is available via: https://www.chemicalsfinder.com/. It is used in the project.

The dataset can be used in the provided format without any conversion needed.

We will keep a copy of the dataset and make it available with our results for the reproducibility.

We will use the dataset as follows: It is a database of chemicals and will be used to identify vendors and order relevant compounds for experimental testing.

- European Lead Factor ESCALATE4COV
  It is available via: https://swissmodel.expasy.org/repository/species/2697049. It is used in the project.

  Owner of this dataset: ESCALATE4COV consortium members.

  The dataset can be used in the provided format without any conversion needed.

  The original dataset will be available both from the provider and from us together with our results for the reproducibility.

  We will use the dataset as follows: The results from the experiments will be used to drive future experiments in the project.

We have found the following non-reference datasets that we have considered for re-use:

- ESCALATE4COV Models
  It is available via: https://swissmodel.expasy.org/repository/species/2697049. It is used in the project.

  Owner of this dataset: ESCALATE4CoV consortium members. The owners of the dataset will collaborate on this project.

  The dataset can be used in the provided format without any conversion needed.

  We will use its online version without downloading it.

  It is a fixed dataset, changes will not influence reproducibility of our results.

  Only part of the dataset will be used; any filtering or selection will be well documented.

  We will use the dataset as follows: This dataset will be used for docking studies and progressing of promising compounds further in the pipeline.

Data formats and types

We will be using the following data formats and types:

- **Extensible Stylesheet Language** (XSL)

It is a standardized format. We are aware that this is not a suitable format for long-term archiving. We expect to have 15 GB of data in this format.

- **Minimum Information for Publication of Quantitative Real-Time PCR Experiments** (MIQE)

  It is a standardized format. We are aware that this is not a suitable format for long-term archiving. We expect to have 15 GB of data in this format.

- **Protein Data Bank Format** (PDB)

  It is a standardized format. This is a suitable format for long-term archiving. We expect to have 200 GB of data in this format.

- **Comma-separated Values** (CSV)

  It is a standardized format. This is a suitable format for long-term archiving. We expect to have 50 GB of data in this format.

- **Document management -- Electronic document file format for long-term preservation -- Part 1: Use of PDF 1.4 (PDF/A-1)** (ISO 19005-1:2005)

  It is a standardized format. This is a suitable format for long-term archiving. We will have only a small amount of data stored in this format.

- **MIAPE: Mass Spectrometry Informatics** (MIAPE-MSI)

  It is a standardized format. This is a suitable format for long-term archiving. We expect to have 200 GB of data in this format.

- **MAGE-TAB for Proteomics Experiments** (MAGE-TAB-Proteomics)

  It is a standardized format. This is a suitable format for long-term archiving. We expect to have 200 GB of data in this format.

- Carl Zeiss Image CZI
  It is a standardized format. This is a suitable format for long-term archiving. We expect to have 50 GB of data in this format.

- Tagged Image File Format (TIFF)
  It is a standardized format. This is a suitable format for long-term archiving. We expect to have 50 GB of data in this format.

- **CSV Dialect Description Format** (CSV-DDF)

  It is a standardized format. This is a suitable format for long-term archiving. We expect to have 100 GB of data in this format.

2. FAIR Data

2.1. Making data findable, including provisions for metadata

- **Viral assay data** (published)
  We will distribute the dataset using:
  - *Domain-specific repository*: **ChEMBL** (ChEMBL)

    We are going to contact the repository.

  - *Special-purpose repository for the project*. We will be able to support this repository for a sufficiently long time. The repository will provide a search and simple access interface.

  There won't be different versions of this data over time.
  We will be adding a reference to the published data to at least one data catalogue.
- **Biochemical assay data** (published)
  We will distribute the dataset using:
  - *Special-purpose repository for the project*. We will be able to support this repository for a sufficiently long time. The repository will provide a search and simple access interface.
- **ADMETox data** (not published)
- **Computational and crystallographic protein models** (published)
  We will distribute the dataset using:
  - *Domain-specific repository*: **RCSB Protein Data Bank** (RCSB PDB)

    We are going to contact the repository.

  - *Domain-specific repository*: **ModelArchive**

    We have already contacted the repository.

- **Machine learning models** (published)
  We will distribute the dataset using:
  - *Special-purpose repository for the project*. We will be able to support this repository for a sufficiently long time. The repository will provide a search and simple access interface.
  - *Domain-specific repository*: **GitHub** (GitHub)

    We don't need to contact the repository because it is a routine for us.

  - *Domain-specific repository*: **Zenodo** (Zenodo)

    We don't need to contact the repository because it is a routine for us.

There will be different versions of this data over time; the versions will be dated and numbered.
We will be adding a reference to the published data to at least one data catalogue.

- **In-vivo experimental data** (published)
  We will distribute the dataset using:
  - *Our national repository:* **Recherche Data Gouv**

There are the following 'Minimal Metadata About ...' (MIA...) standards for our experiments:

- **Minimum Information for Publication of Quantitative Real-Time PCR Experiments** (MIQE)
- **MIAPE: Mass Spectrometry Informatics** (MIAPE-MSI)
- **MAGE-TAB for Proteomics Experiments** (MAGE-TAB-Proteomics)
- **Minimal Information about a high throughput SEQuencing Experiment** (MINSEQE)

We will use an electronic lab notebook to make sure that there is good provenance of the data analysis.

The provenance will be captured using W3C PROV.

We made a SOP (Standard Operating Procedure) for file naming. Files stored on owncloud will be systematically organized by work package names (eg. wpx). within each work package, subdirectories will be created for individual tasks, labeled as task_x.x. all data and analysis files relevant to a specific task will be stored in these corresponding folders. to ensure clarity and seamless integration with electronic lab notebook (eln) entries, subfolders within each task will follow a standardized naming convention: yyyymmdd_elnid_exptype_partner. this format establishes a direct link between experimental entries and their associated data, facilitating easy retrieval and alignment. additionally, provenance information, such as the date and partner name, will be embedded to identify data owners and provide timestamps, enhancing traceability and accountability. for projects on github involving codebases, repositories will adopt a structured format using cookie-cutter templates (https://github.com/cookiecutter/cookiecutter). this ensures consistent and organized repository layouts, promoting better collaboration and reproducibility. We will be keeping the relationships between data clear in the file names. All the metadata in the file names also will be available in the proper metadata.

2.2. Making data accessible

We will be working with the philosophy *as open as possible* for our data.

The data cannot become completely open because of:

- legal reasons

- patent-related business reasons
- we want to publish a paper first

Data will be released only as soon as restrictions are falling away.

Metadata will be openly available including instructions how to get access to the data. Metadata will available in a form that can be harvested and indexed (managed by the used repository / repositories).

We have a consortium agreement that arranges Intellectual Property.

For the reference and non-reference data sets that we reuse, conditions are as follows:

- ChemFinder - It is freely available for any use (public domain or CC0).
- European Lead Factor ESCALATE4COV - It is freely available with obligation to quote the source (e.g. CC-BY).
- ESCALATE4COV Models - It is freely available with obligation to quote the source (e.g. CC-BY).

For our produced data, conditions are as follows:

- **Viral assay data** (published)
  The distributions will be accessible through:
  - *Domain-specific repository*: **ChEMBL**. Open source repository.
  - *Special-purpose repository for the project*. It will be *Shared* with a predefined list of people. We will be able to support this repository for a sufficiently long time. The repository will provice a search and simple access interface.

  A user of this data can use it without any specific software.
  The dataset will published when the project is wrapped up.
- **Biochemical assay data** (published)
  The distributions will be accessible through:
  - *Special-purpose repository for the project*. It will be *Shared* with a predefined list of people. We will be able to support this repository for a sufficiently long time. The repository will provice a search and simple access interface.

  A user of this data can use it without any specific software.
  The dataset will published when the project is wrapped up.
- **ADMETox data** (not published)
- **Computational and crystallographic protein models** (published)
  The distributions will be accessible through:
  - *Domain-specific repository*: **RCSB Protein Data Bank** (RCSB PDB). Open source repository.
  - *Domain-specific repository*: **ModelArchive.** Open source repository.

A user of this data can use it without any specific software.
The dataset will published when the project is wrapped up or a publication is provided.

- **Machine learning models** (published)
  The distributions will be accessible through:
  - *Special-purpose repository for the project*. It will be *Open* (shared with anyone). We will be able to support this repository for a sufficiently long time. The repository will provice a search and simple access interface.
  - *Domain-specific repository*: **GitHub** (GitHub). It will be *Open* (shared with anyone)
  - *Domain-specific repository*: **Zenodo** (Zenodo). It will be *Open* (shared with anyone)

  A user of this data can use it without any specific software.
  The dataset will published when the project is wrapped up.
- **In-vivo experimental data** (published)
  The distributions will be accessible through:
  - *Our national repository:***Recherche Data Gouv.** It will be *Open* (shared with anyone).

  A user of this data can use it without any specific software.
  The dataset will published when the project is wrapped up.

2.3. Making data interoperable

We will be using the following data formats and types:

- **Extensible Stylesheet Language** (XSL)

  It is a standardized format.

- **Minimum Information for Publication of Quantitative Real-Time PCR Experiments** (MIQE)

  It is a standardized format.

- **Protein Data Bank Format** (PDB)

  It is a standardized format.

- **Comma-separated Values** (CSV)

  It is a standardized format.

- **Document management -- Electronic document file format for long-term preservation -- Part 1: Use of PDF 1.4 (PDF/A-1)** (ISO 19005-1:2005)

It is a standardized format.

- **MIAPE: Mass Spectrometry Informatics** (MIAPE-MSI)

  It is a standardized format.

- **MAGE-TAB for Proteomics Experiments** (MAGE-TAB-Proteomics)

  It is a standardized format.

- Carl Zeiss Image CZI
  It is a standardized format.

- Tagged Image File Format (TIFF)
  It is a standardized format.

- **CSV Dialect Description Format** (CSV-DDF)

  It is a standardized format.

We will be using the following standards (encodings, terminologies, vocabularies, ontologies):

- **Protein Data Bank Identifier** (PDB Identifier)
- **Simplified Molecular Input Line Entry Specification Format** (SMILES)

2.4. Increase data re-use

The metadata for our produced data will be kept as follows:

- **Viral assay data** (published) – This data set will be kept available for a fixed period (prepaid) of: 5 years after the project ends.
- **Biochemical assay data** (published) – This data set will be kept available for a fixed period (prepaid) of: 5.
- **ADMETox data** (not published) – This data set will be kept available for a fixed period (prepaid) of: 5 years post project termination.
- **Computational and crystallographic protein models** (published) – This data set will be kept available for a fixed period (prepaid) of: 5 years post project termination. – The metadata will be available even when the data no longer exists.
- **Machine learning models** (published) – This data set will be kept available as long as technically possible. – The metadata will be available even when the data no longer exists.
- **In-vivo experimental data** (published) – This data set will be kept available for a fixed period (prepaid) of: 5 years post project termination. – The metadata will be available even when the data no longer exists.

As explained in Section 2.2, our data cannot become completely open.

There are IP reasons why our data can not be open. It is clear who owns data and documents.

Someone will be given the decision power to move documents or data to a new place after the project has finished.

We will be archiving data (using so-called *cold storage*) for long term preservation already during the project. The data are expected to be still understandable and reusable after a long time.

To validate the integrity of the results, the following will be done:

- We will run a subset of our jobs several times across the different compute infrastructures.
- We will be instrumenting the tools into pipelines and workflows using automated tools.
- We will use independently developed duplicate tools or workflows for critical steps to reduce or eliminate human errors.
- We will run part of the data set repeatedly to catch unexpected changes in results.

3. Other research outputs

We use Data Stewardship Wizard for planning our data management and creating this DMP. The management and planning of other research outputs is done separately and is included as appendix to this DMP. Still, we benefit from data stewardship guidance (e.g. FAIR principles, openness, or security) and it is reflected in our plans with respect to other research outputs.

4. Allocation of resources

FAIR is a central part of our data management; it is considered at every decision in our data management plan. We use the FAIR data process ourselves to make our use of the data as efficient as possible.

- **Antiviral Therapeutics for Rapid Response Against Pandemic Infectious Diseases**

  Following resources will be dedicated to data management and ensuring that data will be FAIR:

  - **Electronic Lab Notebooks (ELN)** - Electronic cataloging of experimental protocols and in-silico approaches.

    The amount is [redacted].

    This resource is allocated for ensuring findability, ensuring interoperability, ensuring reusability, and supporting management of data.

    This cost will be covered by funding grant (grant number: 101137192)

  - **PRISM (GraphPad)** - Comprehensive Analysis and Powerful Statistics toolkit to organize data and perform analysis.

    The amount is [redacted].

    This resource is allocated for ensuring findability, ensuring interoperability, and ensuring reusability of data.

    This cost will be covered by funding grant (grant number: 101137192)

  - **ENOS** - Efficient and adaptable laboratory animal care management software for acquisition of zootechnical data during experiments.

    The amount is [redacted].

    This resource is allocated for ensuring findability and ensuring reusability of data.

    This cost will be covered by funding grant (grant number: 101137192)

We will be archiving data (using so-called 'cold storage') for long term preservation after the project but also already during the project.

None of the used repositories charge for their services.

We have a reserved budget for the time and effort it will take to prepare the data for publication.

Björn Windshügel is responsible for implementing the DMP, and ensuring it is reviewed and revised.

Yojana Gadiya and Leonie von Berlin is responsible for monitoring compliance with DMP during the project lifetime.

Leonie von Berlin is responsible for the management and proficiency of data including data processing, data policies, data guidelines, and data availability.

To execute the DMP, additional specialist expertise is required and we have such trained support staff available.

We require the following hardware or software in addition to what is usually available in the institute: Electronic Lab Notebooks (ELNs) will be implemented as a standard tool for storing, managing, and reporting experimental procedures across collaborating partners. Data analysis and visualization will utilize software such as PRISM, Data Warrior, Spotfire, and IBM SPSS, which will enable efficient data storage, visualization, and statistical analyses. For specialized analytical tasks, Mnova will be employed to process data from NMR studies, while BD FACSuite will be used for analyzing flow cytometry (FACS) data. Molecular dynamic simulations will be performed using AMBERtools and NAMD, and virtual screening experiments will leverage tools such as Autodock Vina, Rbdock, VMD, and LiGen. For structural biology, X-ray diffraction data and crystal structure solutions will be handled with XDS and Aimless (from the CCP4 suite), as well as Phaser. Model building and refinement will be conducted using Phenix and COOT, with Pymol or ChimeraX facilitating the graphical analysis and visualization of protein models. For computational workflows and machine learning model development, platforms such as KNIME, Lexis Platform, and WEKA will be utilized. Zetasizer software will support the measurement of dimensions and Z potential of liposomes, while Celena S will analyze imaging data to investigate the mechanisms of antiviral action. Histological data analysis will be conducted using ZEN Blue Edition and ImageJ, while ENOS will manage zootechnical data. All output from the different computational workflows will be deposited on GitHub for community reuse. Finally, electronic Case Report Forms (eCRFs) will be deployed for the collection of data from phase 1 and phase 2 clinical trials, ensuring streamlined and accurate clinical data management.

5. Data security

Project members will not store data or software on computers in the lab or external hard drives connected to those computers. They will not carry data with them (e.g. on laptops, USB sticks, or other external media). All data centers where project data is stored carry sufficient certifications. All project web services are addressed via secure HTTP (https://owncloud.fraunhofer.de/index.php/f/751243360). Project members have been instructed about both generic and specific risks to the project.

The possible impact to the project or organization if information is lost is small. The risk of information leak in the project or organization is acceptably low. The possible impact to the project or organization if information is vandalised is small.

We are not using any personal information.

The archive will be stored in a remote location to protect the data against disasters. The archive need to be protected against loss or theft. It is clear who has physical access to the archives.

We are running the project in a collaboration between different groups and institutes. A collaboration agreement that describes who can have access to what data in the project is set.

6. Ethics

Data we produce

For the data we produce, the ethical aspects are as follows:

- **Viral assay data**
  - It does not contain personal data.
  - It does not contain sensitive data.
- **Biochemical assay data**
  - It does not contain personal data.
  - It does not contain sensitive data.
- **ADMETox data**
  - It does not contain personal data.
  - It does not contain sensitive data.
- **Computational and crystallographic protein models**
  - It does not contain personal data.
  - It does not contain sensitive data.
- **Machine learning models**
  - It does not contain personal data.
  - It does not contain sensitive data.
- **In-vivo experimental data**
  - It does not contain personal data.
  - It does not contain sensitive data.

Data we collect

We will not collect any data connected to a person, i.e. "personal data".

The data collection is not subject to ethical legislation.

7. Other issues

We use the Data Stewardship Wizard with its *Common DSW Knowledge Model* (ID: dsw:root:2.6.8) knowledge model to make our DMP. More specifically, we use the https://researchers.dsw.elixir-europe.org/wizard DSW instance where the project has direct URL: https://researchers.dsw.elixir-europe.org/wizard/projects/b9597380-87dd-4a50-a2ee-6c636623d3f3.

We will not be using any extra national, funder, sectorial, nor departmental policies or procedures for data management.