Funded by the European Union

AVITHRAPID PROJECT

101137192 - AVITHRAPID

Antiviral Therapeutics for Rapid Response Against Pandemic Infectious Diseases

WP7 FAIR DATA MANAGEMENT

D7.1 Data Management Plan

Lead contributor	1-FRAUNHOFER
Other contributors	All consortium members

Data	**Description**
Due date	30.06.2024
Delivery date	28.06.2024
Delivery type	DMP
Dissemination level	PU

Document History

Version	**Date**	**Description**
V0.1	25-06-2024	First Draft - Yojana Gadiya
V1.0	28-06-2024	Final Draft – Björn Windshügel

Introduction

A Data Management Plan (DMP) is a crucial component of any research project, outlining how data will be handled both during and after the project. It ensures that data is well-organized, accessible, and reusable, enhancing the integrity and reproducibility of research findings. Effective data management is essential for maximizing the impact of research, fostering collaboration, and complying with institutional and funding agency requirements. The principles underlying EU H2020 Programme DMPs (https://ec.europa.eu/research/participants/data/ref/h2020/other/gm/reporting/h2020-tpl-oa-data-mgt-plan-annotated_en.pdf) and the 2023 NIH Data Management and Sharing Policy (https://oir.nih.gov/sourcebook/intramural-program-oversight/intramural-data-sharing/2023-nih-data-management-sharing-policy) have been adopted for the AVITHRAPID DMP .

The AVITHRAPID DMP (Version 1) was written by the Fraunhofer. The AVITHRAPID DMP is a 'living' document that outlines how the project data is handled during and after the project completion, and so it will be regularly evaluated for effectiveness and will be updated, when necessary, by the Fraunhofer. The AVITHRAPID DMP is an official component of the project, and a Deliverable for Work Package 7.

The term "data" does not have one clear definition and can be interpreted differently depending on the context, such as a researcher's field of study. Examples of data are tables of numbers, genomic data, images, survey results, transcripts of interviews and video or audio recordings. Within in the scope of the AVITHRAPID DMP data is defined as "the recorded factual material commonly accepted in the scientific community as of sufficient quality to validate and replicate research findings, regardless of whether the data are used to support scholarly publications".

The AVITHRAPID DMP refers to data related information such as the types of data produced, metadata, policies for access and sharing, and the plans for archiving and preserving data so that they are accessible over time. This will ensure that data will be properly documented and available for use by other researchers in the future. This is accomplished by making data FAIR (findable, accessible, interoperable and re-usable) whenever possible.

The DMP describes the data management life cycle for data from the Project. It covers:

- What type and format of data that will be generated/collected.
- Data handling during and after the Project.
- What methodologies and standards are being applied.
- How the data will be curated and preserved.
- How the data quality for FAIRness will be evaluated.

Finally, the substantial growth in the volume of data being generated across all disciplines requires adequate solutions to be in place for efficient project management. The genuine worth of data lies in its utilization and reutilization and it is expected that innovation driven by data will yield significant benefits for citizens, such as progress in personalized medicine. The collection and utilization of data should align with European values, fundamental rights and established regulations (https://digital-strategy.ec.europa.eu/en/policies/strategy-data).

Methods

The Data Management Plan was drafted using the Data Stewardship Wizard (https://researchers.dsw.elixir-europe.org/wizard) with its Common DSW Knowledge Model (ID: dsw:root:2.6.5) knowledge model. This is a living document with editorial access to project manager (WP8) and WP7 work package leads to allow updation of the DMP as and when required. More

specifically, we use the DSW instance where the project has direct URL: https://researchers.dsw.elixir-europe.org/wizard/projects/b9597380-87dd-4a50-a2ee-6c636623d3f3.

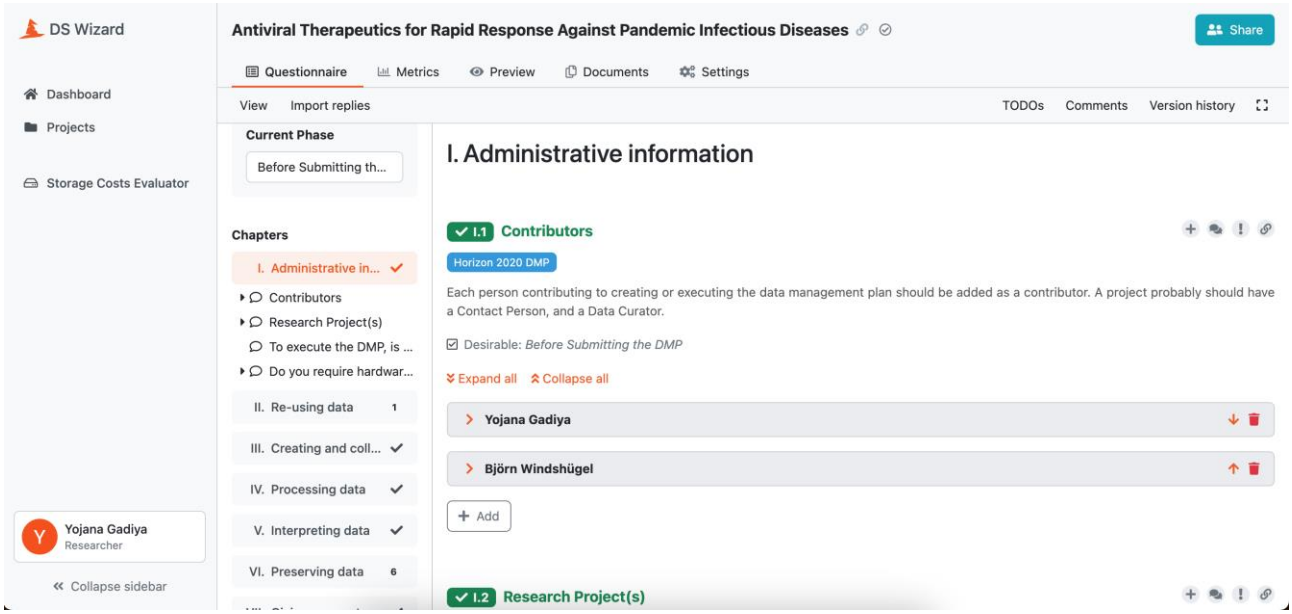

Figure 1: Screenshot of the Data Stewardship Wizard used for drafting the DMP.

To collect the relevant information regarding the data generated across the consortium, a survey based on Excel template was taken. This survery was filled by each partner. The survey capture two broad categories namely:

- Hardware: Capturing all relevant instruments and equipements used by the partner to generate data and relevant metadata.
- Software: Capturing all software solutions that the partner uses to store, query, visualize, and analyze the data.

This document will facilitate the identification of software and hardware solutions used across partners, enabling effective tracking of licenses for each software. Additionally, it will ensure the reproducibility of data by documenting the specific instrumental requirements involved.

Additionally, the main resource for storage of various data types were identified as follow:

- For experimental data and metadata, the Biovia Electronic Lab Notebook (ELN) will be used to record the experiments and data will be depositied in Owncloud following the naming conventioned described later.
- For in-silico analysis and workflows, the scripts will be stored in GitHub and the internal cluster to allow reproducibity and scalibility.

Antiviral Therapeutics for Rapid Response Against Pandemic Infectious Diseases AVITHRAPID

Horizon Europe
Data Management Plan

28 June 2024

History of changes

Version	Date	Description
Draft 1	20.06.2024	Initial draft of DMP
Final version	28.06.2024	Final version

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: Coordinator, Work Package Leader
  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.

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

  These data sets will be collected by expert at european synchotron radiation sources

  The equipment is very well described and known.

  These data are expected to be used by us, as well as deposited on PDB data base and made available to scientific community

- **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.

- **Virtual docking and Protein model experiments**

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

The equipment is very well described and known.

This data are expected to be used only by us.

- **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.

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.

Data formats and types

We will be using the following data formats and types:

- **Extensible Stylesheet Language** (XSL)

  It is a standardized format. This is 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.

2. FAIR Data

2.1. Making data findable, including provisions for metadata

- **Cellular (Antiviral) assay data** (published)
  We will distribute the dataset using:
  - *Domain-specific repository*: **ChEMBL** (ChEMBL); The ChEMBL team will contacted to assist deposition to 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 will be different versions of this data over time; the versions will be dated. 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.

- **ADMET data** (not published)
- **Computational and crystallographic protein models** (published)
  We will distribute the dataset using:
  - *Domain-specific repository*: **RCSB Protein Data Bank** (RCSB PDB); The PDB team will be contacted to assist in data dissemination.

- **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.

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

Our special-purpose repository or project repository will be made accessible via the AVITHRAPID website.

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)

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. The files stored on Owncloud will be organized into work packages (WPX). Within each work package, subdirectories for each task will be created (Task_X.X). All data and analysis files relevant to each task will be stored in these folders. To maintain a clear structure and ensure the linkage between ELN entries and data, subfolders within each task will be named following the format "YYYYMMDD_ELNID_ExpType_Partner". This naming convention will facilitate the association of experimental entries with their corresponding data. Additionally, provenance information, such as the date and partner name, will be included to identify data owners and provide a timestamp. This structured approach will enhance data organization, traceability, and accessibility.

2.2. Making data accessible

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

Not all the data generated within the project will become completely open because of:

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

Data that is not legally restrained will be released after a fixed time period (5 years), unconditionally.

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).

For our produced data, conditions are as follows:

- **Cellular (Antiviral) assay data**
  The distributions will be accessible through **ChEMBL** (https://www.ebi.ac.uk/chembl/) and likewise repositories and through the project repository. 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 project repository.
- **ADMET data** (not published)
- **Computational and crystallographic protein models** (published)
  The distributions will be accessible through **RCSB Protein Data Bank** (RCSB PDB) and likewise repositories.
- **Machine learning models** (published)
  The distributions will be accessible through project repository for prediction on datasets. A user of this models can use it without any specific software.
  The models will published when the project is wrapped up or with a publication.

2.3. Making data interoperable

We will be using the following standard data formats and types:

- **Extensible Stylesheet Language** (XSL)
- **Minimum Information for Publication of Quantitative Real-Time PCR Experiments** (MIQE)
- **Protein Data Bank Format** (PDB)
- **Comma-separated Values** (CSV)
- **Document management -- Electronic document file format for long-term preservation -- Part 1: Use of PDF 1.4 (PDF/A-1)** (ISO 19005-1:2005)
- **MIAPE: Mass Spectrometry Informatics** (MIAPE-MSI)
- **MAGE-TAB for Proteomics Experiments** (MAGE-TAB-Proteomics)

2.4. Increase data re-use

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

- **Cellular (Antiviral) 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)
- **ADMET data** (not published) – This data set will be kept available for a fixed period (prepaid) of: 5 years after project ends.
- **Computational and crystallographic protein models** (published)
- **Machine learning models** (published) – This data set will be kept available as long as technically possible.

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. Making our data FAIR is therefore not a cost that can be separated from the rest of the project.

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.

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 (ELN) will be required to store, manage and report experimental procedures across collaborating partners. Software like PRISIM, Data warrior, Spotfire, and IBM SPSS will be used to store, visualize, and perform statistical analysis on the data. Mnova will be required to analyze data generated from NMR studies and BD FACSuite will be used to analyze FACS data. Molecular dynamic based simulations will be performed using AMBERtools and NAMD, while virtual screening experiments will be run using Autodock Vina, Rbdock, VMD, and LiGen. X-ray diffraction data and crystal structure solution will be done using XDS and Aimless from the CCP4 suite, Phaser, while model building and refinement will be done using Phenix and COOT. Graphical analysis and visualization of protein models will be done using Pymol. Additional software for generation of computation workflows for data analysis and machine learning models involve KNIME, Lexis Platform, and WEKA. Zetasizer Software will be used for measuring the dimension and Z potential of liposomes and Celena S will be used to analyze images for investigating mechanism of antiviral action. eCRF will be used to collect data from clinical trials.

5. Data security

Project members 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 and access controlled HTTP. 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. We will mitigate information leak risk for the project or organization.

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:

- **Cellular (Antiviral) 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.
- **ADMET data**
  - It does not contain personal data.
  - It does not contain sensitive data.
- **Computational 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.

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:2.6.5) 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.