Deliverable 1.5.1

Project ID	654241
Project Title	A comprehensive and standardised e-infrastructure for analysing medical metabolic phenotype data
Project Acronym	PhenoMeNal
Start date of the project	1st September 2015
Duration of the Project	36 Months
Work Package Number	1
Work Package Title	Management
Deliverable Title	D1.5.1 Data Management Plan
Delivery Date	M6
Date of Update	19th July 2016
Work Package leader	EMBL-EBI
Contributing Partners	EMBL-EBI
Authors	Chris Steinbeck, Namrata Kale, Ken Haug, Etienne Thévenot, Philippe Rocca-Serra, Marta Cascante

Abstract: This deliverable describes the initial data management plan for all the submitted and derived research data that will be generated with the PhenoMeNal infrastructure. Each dataset described in this deliverable

includes data set description, standards and metadata, data sharing, ethical and legal compliance and data archiving.
History of Changes: Includes the details about the datasets used in PhenoMeNal in context of description, standards, sharing and archiving.

Table of Contents

1. Executive Summary

The PhenoMeNal project will develop and deploy an integrated, secure, permanent, on-demand service-driven, privacy-compliant and sustainable e-infrastructure for the data processing and analysis pipelines for the molecular phenotype data from the earliest time point of the data acquisition in the laboratory up to the high level medical and biological conclusions and interpretations. It will thus address the challenges arising from extreme data volumes in molecular phenotyping by creating a federated, secure yet high performance e-infrastructure to handle and analyse very large research datasets. To this end, we will use community-accepted open source solutions for analysing metabolomics in conjunction with genomics data, to scale and integrate their approaches and usage into the PhenoMeNal e-infrastructure. The project also aims to provide solutions that bring the compute to the data by providing virtualised compute engines, which can be launched and run on the major available cloud platforms.

In accordance with the H2020 pilot action on open research data, the research data collections assembled as part of the demonstration projects for workflows during the project will be disseminated under a liberal open data license. The open research data will be made freely available, within the appropriate participating data repositories, to the scientific community without restrictions of copyright, patents or other mechanisms of control. However, it should be noted that PhenoMeNal recognises the need for an appropriate balance between openness and confidentiality in the context of handling of sensitive human clinical data. As such, wherever privacy and ethical reasons prevent free data sharing, the issue will be handled in agreement with the national and international laws and regulations for data protection. This is in agreement with Article 29.3 on Open access to research data in the Grant Agreements of the projects according to which:

Regarding the digital research data generated in the action ('data'), the beneficiaries must:

(a) Deposit in a research data repository and take measures to make it possible for third parties to access, mine, exploit, reproduce and disseminate-free of charge for any user-the following:

(i) The data, including associated metadata, needed to validate the results presented in scientific publication as soon as possible;

(ii) Other data, including associated metadata, as specified and within the deadlines laid down in the 'data management plan'

(b) Provide information-via the repository-about tools and instruments at the disposal of the beneficiaries and necessary for validating the results (and-where possible-provide the tools and instruments themselves).

This does not change the obligation to protect results in Article 27, the confidentiality obligations in Article 36, the security obligations in Article 37 or the obligations to protect personal data in Article 39, all of which still apply.

As an exception, the beneficiaries do not have to ensure open access to specific parts of their research data if the achievement of the action's main objective in Annex 1, would be jeopardised by making those specific parts of the research data openly accessible. In this case, the data management plan must contain the reasons for not giving access.

This deliverable describes the initial data management plan for the research data used during the course of the project describing the data sets, standards and metadata, data sharing, ethical and legal compliance and data archiving[1] and will be updated as deliverables D1.5.2 (M18) and D1.5.3 (M30) respectively.

2. Data Sets

PhenoMeNal is an e-infrastructure for managing, preserving and computing with biomedical phenotyping in combination with genomic data that phenome centres and biomedical laboratories will generate from human research subjects. Thus, within the PhenoMeNal project, the research data will essentially be data deposited by the users, a.k.a. "data providers", of the infrastructure, prior to externally generating and producing said data. The project will facilitate the storage of relevant data in secure public environments using data standards and procedures developed by COordination Of Standards In MetabOlomicS (COSMOS, http://www.cosmos-fp7.eu/), the Metabolomics Standards Initiative (MSI, http://www.metabolomics-msi.org/) and European Translational Information & Knowledge Management Services (eTRIKS, https://www.etriks.org/). Metagenomic, genomic and metabolomics data and protocols will, where appropriate, be deposited with the European Bioinformatics Institute (EMBL-EBI, UK). We also have the ambitions to in the future include the Mosler secure computing environment (https://bils.se/resources/mosler.html) running in Uppsala, which is governed by similar privacy protection mechanisms as the EMBL-EBI EGA (European Genome-phenome archive).

2.1. PhenoMeNal use case datasets

The PhenoMeNal project does not generate novel data, but in order to properly develop the software environment, a number of standard datasets will be used to test the formats, data processing pipelines, user interaction and the stability of the software and to make sure that our procedures are in line with generally accepted Ethical Legal and Social Implications (ELSI) guidelines. For this reason, we have selected a number of use cases that typically reflect the type of data that will be used in PhenoMeNal.

2.1.1. Restricted datasets

Data set descriptions

The initial use of restricted data will be for development and testing of software and computational tools, but no biological analysis will be performed or published.

2.1.1.1.The MESA dataset

The Multi-Ethnic Study of Atherosclerosis (MESA, http://www.mesa-nhlbi.org/) is a medical research study involving more than 6,000 men and women in the United States. The study focuses on the characteristics of subclinical cardiovascular diseases. As part of the COMBI-BIO project (Development of combinational biomarkers for subclinical atherosclerosis, http://www.combi-bio.eu/), metabolomics data was produced in two phases for 4,000 MESA participants from serum samples using NMR and LC-MS platforms.

2.1.1.2. The CoLaus dataset

The main goals of this study are to obtain information on the epidemiology and genetic determinants of cardiovascular risk factors and diseases as well as mental health in the adult population of Lausanne.

2.1.1.3.Uppsala Fibromyalgia study

Metabolite-profiling data from a cohort of 120 participants, consisting of fibromyalgia patients and several matched controls.

Standards and metadata

PhenoMeNal will not impose any additional metadata requirements for privacy-restricted datasets. These datasets are governed by their existing ELSI requirements, and will have been defined during the design of each study respectively.

Data sharing

The target audience for these restricted datasets are the researchers within the PhenoMeNal consortium and none of this data will be made publicly available. The use of the data will be behind secure firewalls and used for testing purposes of the software processing and analysis. We do not expect to publish any biological findings.

Ethical and legal compliance

We have ethical approval, with accompanying documentation detailing consent information.

Archiving and preservation

The data will be deposited in the secure EGA database (http://www.ebi.ac.uk/ega/) at the EMBL-EBI. The data will not be public and will be accessible, through normal EGA data access procedures, only to the members of the PhenoMeNal consortium for testing purposes. PhenoMeNal and EGA will not impose additional metadata requirements for these datasets, in accordance with existing practices of the EGA.

2.1.2. Open access datasets

Data set descriptions

These datasets are currently available as public or pre-publication datasets in the MetaboLights (http://www.ebi.ac.uk/metabolights/) repository and will also be used as use cases for testing of software components. These datasets already contain metadata according to the metadata standards developed by the EMBL-EBI lead initiatives like COSMOS FP7 (http://www.cosmos-fp7.eu/) and the MetaboLights project (BBSRC). MetaboLights require all datasets to be MSI compliant (http://www.metabolomics-msi.org) and annotated using the ISA-Tab format (http://isa-tools.org/format/specification/). All data in MetaboLights is governed by EMBL-EBI terms of use http://www.ebi.ac.uk/about/terms-of-use.

2.1.2.1. “The *sacurine* dataset: Physiological Variations of the Urine Metabolome”

To determine the variations of the urine metabolome with age, body mass index, and gender, under basal (i.e., physiological) conditions, urine samples from a cohort of 183 human adults have been analysed by liquid chromatography coupled to high-resolution mass spectrometry[2].

Within PhenoMeNal, the objectives are:

a. To reproduce the publicly available reference workflow {(http://workflow4metabolomics.org/dataset_sacurine; Workflow4metabolomics (http://workflow4metabolomics.org) infrastructure} on the PhenoMeNal cloud environment,

b. To facilitate the standardization of the metadata by using PhenoMeNal guidelines and modules.

Standards and metadata

Metabolites are annotated according to the guidelines from the Metabolomics Standards Initiative[3].

Data sharing

Raw data are publicly available on the Workflow4metabolomics infrastructure. They will also be made available in the EMBL-EBI MetaboLights repository during the PhenoMeNal project.

Ethical and legal compliance

This is an open data publicly available, anonymised and filtered. It does not include consent forms, ethical approval or patient information at this source.

Archiving and preservation

The data is currently archived in the Workflow4metabolomics infrastructure and the MetaboLights database, and is publicly available. Workflow4metabolomics is funded by two French government infrastructures in the medium term (MetaboHUB, http://www.metabohub.fr: National Infrastructure for Metabolomics and Fluxomics; and IFB, http://www.france-bioinformatique.fr: French Bioinformatics Institute). In addition, in accordance with EMBL-EBI policy, the operation and running of the strategic MetaboLights archive is centrally funded and will be maintained without the need for short or medium term funding.

2.1.2.1 Data from fluxomic analysis

Determination of metabolic flux distributions is fundamental in order to have a complete characterization of metabolic phenotypes. In these analyses, cells are incubated in the presence of isotope-enriched substrates to decipher their biochemical processing through the main metabolic pathways. This type of analyses is not normally applied in-vivo, but it can be done in primary cultures isolated from patients or as recently demonstrated, in ex-vivo tissue slices[4]. The use of cells in primary culture from patients is a promising tool looking to evaluate the differential response of control and patient cells to drugs, including anticancer agents. There are currently no cohort studies covering patient samples. However, a dozen datasets corresponding to cancer focused, ex-vivo and in-vitro studies using mass spectrometry based stable isotope resolved metabolomics, are available from EMBL-EBI MetaboLights repository or NIH Metabolomics Workbench (http://www.metabolomicsworkbench.org/nihmetabolomics/). These datasets are already scheduled for publication during the course of the project, and as such will be used as examples for our analysis capabilities. As data owners are existing PhenoMeNal consortia members, data accessibility will be unhindered.

Standards and metadata
These datasets are annotated according to the current metadata requirements of EMBL-EBI MetaboLights, including raw data for reuse and reproducibility.

Data sharing
The data is currently under submitter embargo in pre-publication status and will be made publicly available during the course of the project.

Ethical and legal compliance
Upon reaching the submitters embargo date, this open data will be publicly available, and is already anonymised and filtered. It does not include consent forms, ethical approval or patient information and is governed by EMBL-EBI terms of use: http://www.ebi.ac.uk/about/terms-of-use.

Archiving and preservation
The data is currently archived in the MetaboLights database and is publicly available. In accordance with EMBL policy, the operation and running of this strategic archive is centrally funded and will is maintained without the need for short or medium term funding.

2.2. MetaboLights data repository

The MetaboLights repository is hosted at the EMBL-EBI and is a database for metabolomics experiments and derived information. MetaboLights accepts and stores all types of metabolomics data. According to EMBL-EBI terms of use, all public datasets are open and available for any purpose.

Data set descriptions

MetaboLights includes datasets submitted by the metabolomics user community worldwide and are cross-species and cross technique. The types of data include experimental NMR, LC-MS, GC-MS, Imaging and chromatographic data.

Standards and metadata

The MetaboLights submission pipeline is utilising the ISA software suite (http://isa-tools.org/). All experimental data is extensively annotated in ISA-Tab format. MetaboLights enforces rigorous annotation requirements, set out in the MSI recommendations. Additionally MetaboLights requirements for both raw and open source data formats ensure that the primary research data is easily reusable.

Data sharing

The datasets within Metabolights are archived either as pre-publication (private accessible only to the submitter) or as public datasets. As of summer 2016 MetaboLights holds over 100 human datasets, of which about 60% is in the public domain.

Ethical and legal compliance

According to MetaboLights guidelines submitters required to anonymised and pre-filter all datasets prior to submission to the archive. Submissions do not include consent forms, ethical approval or patient information and is governed by EMBL-EBI terms of use: http://www.ebi.ac.uk/about/terms-of-use.

Archiving and preservation

All data in the MetaboLights database is publicly available after curation approval and reaching the submitters embargo date. In accordance with EMBL policy, the daily operation and running of this strategic archive is centrally funded and is therefore maintained without the need for short to medium term funding.

2.3. VRE data

The PhenoMeNal VRE will facilitate the analysis of private and public human molecular phenotyping data and metadata through virtualised workflows, upholding privacy and ethical requirements by enabling compute to the data by running PhenoMeNal virtual machines locally. Users can use public or private (personal) datasets for performing online analysis. It should be noted that PhenoMeNal will not support permanent direct data sharing from within a VRE. For sharing the datasets will have to be migrated to public repositories linked to PhenoMeNal pipeline, such as MetaboLights, and the data will consequently be governed by the respective repository data management policies, audit policies and data submitter embargo periods.

2.3.1.Public Test VRE

PhenoMeNal will offer a public test version of the VRE, purposed for testing the integrated tools and workflows. Users can register and upload data for the sole purpose of testing the tools and workflows contained within. However this is not for the purpose of persisting or sharing the files or the derived information. The test VRE will be subjected to regular rebuilds, hence data will be removed. No sensitive data should be submitted to the test VRE. This will be further detailed in the terms and conditions for the VRE, which users are bound to accept upon creating of a user account.

Standards and Metadata
In this VRE the only constraint is the format of the raw or open source data files. The individual tools either contained in a workflow or running independently, will have different requirements for what type of data files can be processed. Each tool and compatible file formats will be detailed in the VRE App Library.

Data sharing
No data will be shared from within the test VRE.

Ethical and legal compliance
No sensitive or private data should be uploaded to the test VRE. This will be further detailed in the terms and conditions for the VRE.

Archiving and preservation
No data will be preserved long term in the test VRE. Migrating data to a public repository will not be enabled from the test VRE.

2.3.2.User controlled VRE

Users creating a personal VRE on the PhenoMeNal infrastructure, or with a supported public cloud supplier, are able to control the data therein. A personal VRE enables extended data upload and capture of derived data. In this VRE the only constraint is the format of the raw or open source data files. The individual tools, either contained in a workflow or running independently, will have different requirements for what type of datafiles can be processed. Each tool and compatible file formats will be detailed in the VRE App Library.

Data uploaded for analysis and the resulting data (derived) from this process is naturally available to the user. Tenancies within the PhenoMeNal infrastructure will have to be time limited to enable fair sharing of available resources. This is obviously not the case where the user has deployed the VRE to a commercial cloud provider. Here the user is only restricted to their financial contract with the commercial supplier.

2.3.2.1. Public data

Published data where the data owners have already ensured that they have sought and obtained all appropriate approvals, ethical and legal, for the data collected, clearly simplifies where data is processed and later published. This section details the plans for handling data with all existing ELSI related approvals.

Standards and Metadata
PhenoMeNal will not impose any additional requirements for metadata for datasets. These datasets are governed by their existing privacy and ethical requirements, and metadata requirements will have been defined during the design of each study.

Data sharing
We will not offer direct data sharing from within the VRE, however we will facilitate tools and mechanisms to publish the uploaded and/or derived data to the participating data repositories. Guidance will be provided for data depositions into the existing public repositories linked to PhenoMeNal pipeline, MetaboLights, and will be governed by repositories data management policies, audit policies and data submitter embargo periods.

Ethical and legal compliance
It is the sole responsibility of the data provider to ensure that they have sought and obtained the data in compliance with all ethical and legal approvals. Use of identifiable data after consent only and use of anonymised data whenever possible. This will be governed by PhenoMeNal terms of use. See Annex 3.1

Archiving and preservation
Beyond this scope and will be governed by repositories data management policies, audit policies and data submitter embargo periods.

2.3.2.2. Restricted data

Data that is governed by existing privacy and ethical requirements can only be uploaded and used in this VRE when explicit has been granted. Users of the VRE will have to ensure this permission has been granted prior to uploading any data. Where such permission is not in place, the user should rather deploy the VRE into their own controlled infrastructure.

Standards and Metadata
PhenoMeNal will not impose any additional requirements for metadata for datasets. These datasets are governed by their existing privacy and ethical requirements, and metadata requirements will have been defined during the design of each study.

Data sharing
We will not offer direct data sharing from within the VRE, however we will facilitate tools and mechanisms to publish the uploaded and/or derived data to the participating data repositories. Guidance will be provided for data depositions into the existing public repositories linked to PhenoMeNal pipeline, EMBL-EBI EGA (http://www.ebi.ac.uk/ega/), and will be governed by repositories data management policies, audit policies and data submitter embargo periods.

Ethical and legal compliance
It is the sole responsibility of the data provider to ensure that they have sought and obtained the data in compliance with all ethical and legal approvals. Use of identifiable data after consent only and use of anonymised data whenever possible. This will be governed by PhenoMeNal terms of use. See Annex 3.1

3. Annexes

3.1. Phenomenal Terms of Use version 1.0

PhenoMeNal is an integrated, secure, on-demand service-driven, privacy-compliant and sustainable European e-infrastructure for processing, analysis and information mining of metabolomics data. The project has been designed to enable maximum benefit from research by making data as accessible as possible to the research community, while protecting the interests of participants from whom the data originate with regard to Ethical, Legal and Social Implications (ELSI) and within the scope of their consent. These Terms of Use reflects PhenoMeNal's commitment to provide this service and impose no additional constraints on the use and transfer of the contributed data than those provided by the data owner.

- All users have an obligation of confidentiality and must conform to data protection principles to ensure that data is processed in compliance with the legal and ethical requirements.
- The data owners must ensure that they have sought and obtained, where necessary, all appropriate approvals, ethical and legal, for the data collected.
- For animal data, the data owner must ensure that national guidelines for their welfare and care during the collection of data have been followed.
- PhenoMeNal does not guarantee the accuracy of any provided data.
- PhenoMeNal has implemented appropriate technical and organisational measures to ensure a level of security which we deem appropriate, taking into account the sensitivity of data we handle. However, the data provider holds sole responsibility for the usage and distribution of data.
- Computing of personal and sensitive data on PhenoMeNal infrastructure should be run internally by the users on their secure cloud infrastructures under appropriate firewalls. PhenoMeNal will not hold any liability for any loss or damage to data.
- While we will retain our commitment to privacy of sensitive data, we reserve the right to update these Terms of Use at any time. When alterations are inevitable, we will attempt to give reasonable notice of any changes by placing a notice on our website, but you may wish to check each time you use the website. The date of the most recent revision will appear on this, the 'PhenoMeNal's Terms of Use' page. If you do not agree to these changes, please do not continue to use our

services. We will also make available an archived copy of the previous Terms of Use for comparison.

- Any questions or comments concerning these Terms of Use can be addressed to: phenomenal-help@ebi.ac.uk

4. References

1. Guidelines on Data Management plan in Horizon 2020 http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf
2. Thévenot, E.A., Roux, A., Xu, Y., Ezan, E. and Junot, C. (2015). Analysis of the Human Adult Urinary Metabolome Variations with Age, Body Mass Index, and Gender by Implementing a Comprehensive Workflow for Univariate and OPLS Statistical Analyses. J Proteome Res.;14(8): 3322-35.
3. Roux, A., Xu, Y., Heilier, J.F., Olivier, M.F., Ezan, E., Tabet, J.C. and Junot, C. (2012) Annotation of the human adult urinary metabolome and metabolite identification using ultra high performance liquid chromatography coupled to a linear quadrupole ion trap-Orbitrap mass spectrometer. *Anal Chem*. 84(15): 6429-37.
4. Fan, T.W., Lane, A.N. and Higashi, R.M. (2016) Stable Isotope Resolved Metabolomics Studies in Ex Vivo TIssue Slices. *Bio Protoc*. 6(3). pii: e1730.