

Deliverable 1.5.2

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	D1.5.2 Updated Data Management Plan
Date of update	10th August 2017
Version	2.1
Delivery Date	M18
Work Package leader	EMBL-EBI
Contributing Partners	EMBL-EBI, UOXF, UU, CEA, IPB, UB, ICL
Authors	Namrata Kale, Ken Haug, Philippe Rocca-Serra, Kim Kultima, Etienne Thevenot, Daniel Schober



Abstract: This deliverable is the updated 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

Changes from D1.5.1 - 15.02.2017:

- It was concluded that Mosler cannot be used due to technical restrictions. All references to Mosler are removed
- URLs added as footnotes for readability
- Added clarification on FAIR data
- Added information on metabolomics study size
- Added information on data security
- Appropriate consent forms to use CoLaus dataset could not be obtained. All references regarding the dataset as a use case are removed
- Added information on the data produced/generated during the project

Changes from D1.5.2 - 10.00.2017:

- Added description of non-scientific data, a.k.a. "Other Data"



1	EXECUTIVE SUMMARY	5
2	DATA SUMMARY	6
2.1	Scientific data	7
2.1.1	PhenoMeNal use case datasets	7
2.1.1.1	Restricted datasets	7
2.1.1.2	Open access datasets	9
2.1.2	MetaboLights data repository	12
2.1.3	VRE data	13
2.1.3.1	Public Test VRE	13
2.1.3.2	User controlled VRE	14
2.2	Other data	16
2.2.1	PhenoMeNal administrative data	16
2.2.1.1	User accounts	16
2.2.1.2	Email distribution lists	17
2.2.1.3	Online document library and working area	17
2.2.1.4	Online meetings and Instant messaging	17
2.2.1.5	Project Management	18
2.2.2	Software, tools and website	18
2.2.2.1	Source code	18
2.2.2.2	Support tickets	18
2.2.2.3	WordPress backups	18
2.2.2.4	Continuous integration	19
2.2.3	Dissemination material	19
3	ETHICAL ASPECTS	19
3.1	PhenoMeNal Data managers and Data controller	19
3.2	Data Availability process	20
4	DATA SECURITY	20
4.1	Cloud tenancy	20
4.2	Data preservation	20
4.2.1	Data embargo and secure transfer	20
5	ANNEXES	22
5.1	Renewed permission to use MESA dataset	22



5.2	Phenomenal Terms of Use version 1.0	23
5.3	Guidance for Data Governance Managers	24
	List of Data Governance Managers in PhenoMeNal	25
6	References	27



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 and will be updated next as deliverables D1.5.3 (M30) respectively.

2 DATA SUMMARY

PhenoMeNal is an e-infrastructure for managing, preserving and computing with biomedical phenotyping in combination with genomic data that phenome centers 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 **CO**ordination Of **S**tandards In **MetabO**lomic**S** (COSMOS¹), the Metabolomics Standards Initiative (MSI²) and European Translational Information & Knowledge Management Services (eTRIKS³). Metagenomic, genomic, metabolomics data and protocols will, where appropriate, be deposited with the European Bioinformatics Institute (EMBL-EBI, UK).

EMBL-EBI MetaboLights⁴, is the integrated data repository for PhenoMeNal. The average study collection size in MetaboLights is about 25Gb⁵, however the median size is around 1.25Gb per study. Due to the homogeneity of metabolomics data, MetaboLights accepts all types of metabolomics data from the worldwide metabolomics

¹ <http://www.cosmos-fp7.eu/>

² <http://www.metabolomics-msi.org/>

³ <https://www.etriks.org/>

⁴ <http://www.ebi.ac.uk/metabolights/>

⁵ <http://www.ebi.ac.uk/metabolights/statistics>



community. All public data in MetaboLights is freely available by any individual and for any purpose, as per EMBL-EBI guidelines⁶. Our previous ambition of using the Nordic Mosler system, was revised after initial conversations via UU. It was concluded that Mosler could not be used due to technical restrictions⁷. In PhenoMeNal useable data types are very much dependent on the capabilities of the chosen workflow and/or tools therein. We have endeavored to describe these capabilities in the technical description available through the service catalog⁸ for each tool we support.

2.1 Scientific data

2.1.1 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.

2.1.1.1 Restricted datasets

Data set descriptions

The PhenoMeNal consortium initially selected 3 datasets as its use cases. These are a) MESA b) Fibromyalgia c) CoLaus. The use of these datasets will only be for development and testing of software and computational tools, but no biological analysis will be performed or published.

The ethical approvals for the MESA were renewed. However, The CoLause dataset was withdrawn due to lack of appropriate data consents while the Fibromyalgia dataset is now restricted to PhenoMeNal researchers from Akademiska Sjukhuset⁹ only.

A. *The MESA dataset*

The Multi-Ethnic Study of Atherosclerosis, MESA¹⁰, 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

⁶ <http://www.ebi.ac.uk/about/terms-of-use>

⁷ <https://bils.se/resources/mosler.html>

⁸ <http://portal.phenomenal-h2020.eu/app-library>

⁹ <http://www.akademiska.se/en/>

¹⁰ <http://www.mesa-nhlbi.org/>



project (Development of combinational biomarkers for subclinical atherosclerosis¹¹), metabolomics data was produced in two phases for 4,000 MESA participants from serum samples using NMR and LC-MS platforms.

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. Recently, the ethical approvals for the use of dataset was renewed. See Annex 5.1.

Archiving and preservation

The data will be deposited in the secure European Genome-phenome Archive (EGA database)¹² 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.

B. 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.

¹¹ <http://www.combi-bio.eu/>

¹² <http://www.ebi.ac.uk/ega/>



Data sharing

The target audience for these restricted datasets are the PhenoMeNal researchers within Akademiska Sjukhuset for testing 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 is accessible for PhenoMeNal researchers within Akademiska Sjukhuset. The data will not be public and access will only be given to the members of the PhenoMeNal consortium for testing purposes within Akademiska Sjukhuset.

2.1.1.2 Open access datasets

Data set descriptions

These datasets are currently available as public or pre-publication datasets in the MetaboLights repository and will also be used as use cases for testing of software components in PhenoMeNal. These datasets already contain metadata according to the metadata standards developed by the EMBL-EBI led initiatives like COSMOS FP7 and the MetaboLights project (BBSRC). MetaboLights require all datasets to be MSI compliant and annotated using the ISA-Tab format¹³. The ISA format, endorsed by several data repositories (MetaboLights, GigaDB¹⁴, Dryad¹⁵) and publishers (Nature Publishing Group, Oxford University Press) is now supported by an API¹⁶ and visualisation/search tool¹⁷. The ISA metadata format, and the MSI approved open access raw data formats (mzML and nmrML) used under MetaboLights and PhenoMeNal are one central component required to realise the vision of making data 'Findable, Accessible, Identifiable and Reusable' (FAIR). Recent efforts focused on making data 'Findable', by mapping ISA element to schema.org entities (schema.org is used by search engine such as Bing, Google or Yandex for indexing) and 'Reusable', mainly by implementing new ISA configurations and validation rules, which will result in regularised and augmented semantic markup. This latter step will also increase the

¹³ <http://isa-tools.org/format/specification/>

¹⁴ <http://gigadb.org>

¹⁵ <http://datadryad.org>

¹⁶ <https://github.com/ISA-tools/isa-api>

¹⁷ <http://isa-explorer.org>



number of entities made identifiable and therefore linkable. In a broader context, it means that metabolomics data becomes more discoverable. This is a positive trend to facilitate data reuse, be it for quality assessment, citation and impact rating or for aggregation in large meta-analysis projects. All data in MetaboLights is governed by EMBL-EBI terms of use.

A. 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 184 human adults have been analyzed by liquid chromatography coupled to high-resolution mass spectrometry¹⁸.

Within PhenoMeNal, the objectives are:

- To reproduce the publicly available reference workflow (W4M00001_Sacurine-statistics¹⁹; Workflow4Metabolomics²⁰ e-infrastructure) on the PhenoMeNal cloud environment,
- To facilitate the standardisation of the metadata by using PhenoMeNal guidelines and modules.

Standards and metadata

Metabolites are annotated with identification evidence levels according to the guidelines from the Metabolomics Standards Initiative.

Data sharing

Raw data are publicly available on the Workflow4Metabolomics e-infrastructure. Within the PhenoMeNal project, the full data set was also made available in the EMBL-EBI MetaboLights repository as study MTBLS404²¹.

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

¹⁸ [Thévenot et al, 2015, http://pubs.acs.org/doi/full/10.1021/acs.jproteome.5b00354](http://pubs.acs.org/doi/full/10.1021/acs.jproteome.5b00354)

¹⁹ <http://dx.doi.org/10.15454/1.4811121736910142E12>

²⁰ <http://workflow4metabolomics.org/>

²¹ <http://www.ebi.ac.uk/metabolights/MTBLS404>



two French government infrastructures in the long term (MetaboHUB²²: National Infrastructure for Metabolomics and Fluxomics; and IFB²³: French Institute of Bioinformatics - Elixir node). In addition, in accordance with EMBL 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.

B. Data from fluxomic analysis

Determination of metabolic flux distributions is fundamental in order to have a complete characterisation 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 slices²⁴. 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, have been deposited to EMBL-EBI MetaboLights repository and NIH Metabolomics Workbench²⁵. These datasets are 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.

²² <http://www.metabohub.fr/home.html>

²³ <http://www.france-bioinformatique.fr/en>

²⁴ Fan, T. W., Lane, A. N., and Higashi, R. M. (2016). Stable Isotope Resolved Metabolomics Studies in Ex Vivo Tissue Slices. *Bio Protoc.* 2016 Feb 5;6(3). pii: e1730).

<http://www.ncbi.nlm.nih.gov/pubmed/27158639#>

²⁵ <http://www.metabolomicsworkbench.org>



Ethical and legal compliance

Upon reaching the submitters embargo date, this open data will be publicly available, and is already de-anonymised and filtered. It does not include consent forms, ethical approval or patient information and is governed by EMBL-EBI terms of use.

Archiving and preservation

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

2.1.2 MetaboLights data repository

The MetaboLights repository is hosted at the EMBL-EBI and is a database for metabolomics experiments and derived information. MetaboLights accept and store 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 technique²⁶. 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 suite²⁷. 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.

²⁶ <http://www.ebi.ac.uk/metabolights/studies>

²⁷ <http://isa-tools.org/>



Ethical and legal compliance

According to MetaboLights guidelines submitters are required to de-anonymise 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.

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.

2.1.3 VRE data

The PhenoMeNal VRE²⁸ facilitates analysis of human molecular phenotyping data and metadata through virtualised workflows. Privacy and ethical requirements for private data are ensured by running PhenoMeNal virtual machines locally, behind clinical firewall, exemplifying the 'bring-the-compute-to-the-data' approach. 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 data will have to be migrated to public repository linked to PhenoMeNal the 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. Data generated or uploaded to a VRE remains in the VRE until that user actively removes it or exports it elsewhere. A data retention plan is currently being produced to ensure transparency in duration and how data is retained. PhenoMeNal as such does not actively move data around or outside the VRE, although it provides users the ability to do so. User controlled personal VREs where the owner removes their own instance, will naturally not retain any data.

2.1.3.1 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.

²⁸ <http://portal.phenomenal-h2020.eu/home>



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.

2.1.3.2 User controlled VRE

Users creating a personal VRE, or personal Galaxy²⁹ account, 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 data files 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.

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.

²⁹ <http://public.phenomenal-h2020.eu>



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 repository linked to PhenoMeNal pipeline, MetaboLights, and will be governed by existing repository 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 is governed by PhenoMeNal terms of use³⁰. See Annex 5.2

Archiving and preservation

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

Restricted data

Data that is governed by existing privacy and ethical requirements can only be uploaded and used in this VRE. 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.

³⁰ <http://phenomenal-h2020.eu/home/wp-content/uploads/2016/09/Phenomenal-Terms-of-Use-version-11.pdf>



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 repository linked to PhenoMeNal pipeline, EMBL-EBI 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.

2.2 Other data

In addition to the more scientific data described above, the project generates non-scientific/other types of data.

2.2.1 PhenoMeNal administrative data

Administrative data is data that is related to the execution of workflows and operation of the websites, in addition to user interactions.

2.2.1.1 User accounts

Public accounts

PhenoMeNal has co-developed a new single sign-on (SSO) system³¹ with Elixir. This SSO solution enable users to register and log in to the PhenoMeNal VRE portal using authentication from their personal Google, LinkedIn, ORCID or Edugain institutional accounts. The Elixir SSO retains the user name and email details, and the information is stored in conjunction with a new Elixir person id³², the user's group entitlement (member, administrator etc) within the SSO system and a unique generated id key. The password information is protected and is not retained. Each PhenoMeNal VRE user will have this joint SSO user account. During the signup process, the user will also create a personal PhenoMeNal Galaxy workflow and Jupyter user account. This account is unique to the PhenoMeNal public or personal VRE(s).

When a VRE is deployed to a public/commercial cloud provider (AWS, GCP or OpenStack) the user provides the credentials required by the respective infrastructure,

³¹ <https://www.elixir-europe.org/services/compute/aai>

³² username@elixir-europe.org



like access keys. This information is used to create and remove (destroy) deployments and is only temporarily stored for this purpose.

In a deployed VRE, ie. including deployed within local institutional firewalls, Elixir PhenoMeNal SSO is not required, only the local Galaxy/Jupyter user accounts are used. The Elixir SSO is only required in the PhenoMeNal portal, used to deploy VREs on behalf of the users. Local server logins will be managed by the local institute.

Infrastructure administrative accounts

Users of commercial, or open academic, cloud providers will need to have access to their personal accounts to deploy the PhenoMeNal infrastructure. This information is not known, nor will it have to be known, in the context of the PhenoMeNal project.

The publicly available PhenoMeNal VRE has a set of administrative accounts, like Linux and server logins. These are not known outside a few select project members. Administration of our project WordPress site is managed likewise.

2.2.1.2 Email distribution lists

At this stage in the project, we only operate project internal email distribution lists (DL). The main DL is hosted on Google Groups and all project members are included. This DL is not publicly available and only members can browse the archives.

2.2.1.3 Online document library and working area

In addition to the Google Group described above, the PhenoMeNal project greatly relies on Google Drive for our document archive. The private Google Drive is the only recognised document library for the project and is accessible to all the members of the consortium. There is no general public access to this drive, but select general documents are shared in public read-only mode. This drive is automatically backed up on several computers in real time as most members use the provided Google Drive Sync software. EMBL-EBI also actively use a private Google Suite for privileged access to project management related information. Official project deliverables are publicly available on Zenodo³³, in addition to the project website³⁴, as agreed with the commission.

2.2.1.4 Online meetings and Instant messaging

To ensure continuous effective communication the project frequently use Google Hangout for video conferencing. This is hosted through EMBL-EBI's private Google Suite ensuring only invited project members can join. In addition to the extra security for

³³ <https://zenodo.org/communities/phenomenalh2020/>

³⁴ <http://phenomenal-h2020.eu/home/about/objectives/>



Google Suite Hangouts, we are also able to host 25 concurrent connections, as opposed to the normal 15 for the public free Hangout accounts.

The project relies on a private Slack project³⁵ for the majority of direct, one-to-one or group, communication. This greatly reduces the volume of emails in the project and ensures effective discussions.

2.2.1.5 Project Management

PhenoMeNal is using Pivotal Tracker³⁶ for tracking tasks and deadlines for deliverables. Each task is assigned to individuals and/or an institute. Pivotal Tracker is free for non-profits and academic institutes. Each project member has a user account, linked to the same email as used in the Google Groups email distribution list. All tasks are retained after completion and used in project reporting.

2.2.2 Software, tools and website

2.2.2.1 Source code

All source code generated in the project is open source and stored in GitHub. We have created a project organisation in GitHub³⁷ where all our sub-repositories are stored. All code generated in this project is diligently uploaded and made publicly available under this GitHub structure. All our technical documentation the project Wiki is also facilitated using GitHub.

2.2.2.2 Support tickets

On the main project website, we have set up a helpdesk function, using a WordPress application (not developed by the project). Users, on anyone interested in the project, can ask questions and/or create a support request (ticket). These tickets are triaged by a select few project members and allocated to the appropriate person to be resolved. After the user has asked a question using the online forms, all consecutive communication can be handled using email. Information from the forms and resulting emails are stored in the local WordPress database.

2.2.2.3 WordPress backups

The project website is built using WordPress. All pages and associated data is backed up to your general Google Drive for safe keeping.

³⁵ <https://phenomenal.slack.com>

³⁶ <http://phenomenal-h2020.eu/home/about/objectives/project-management-tool/>

³⁷ <https://github.com/phnmnl>



2.2.2.4 Continuous integration

We use a continuous integration (CI) system³⁸ to build all the source code and containers. This is referencing the repositories under our public GitHub PhenoMeNal organisation. Containers are built by the CI whenever there is a change in a container's GitHub repository source code. Successful container building, which includes tests of the container, triggers the execution of further tests with real data, replicating real production conditions. The definition and setup of this CI system is backed up, no further data is generated during this nightly build process.

2.2.3 Dissemination material

The project dissemination material generated in the form of project reports, deliverables submitted to the commission, flyers, public presentations, posters, videos, user documentation and training material are/will be available on the project website. In addition these are/will be made available on Zenodo as open access under creative commons attribution. All scientific publications are open access and are accessible via OpenAire³⁹. Video demonstrations available on YouTube are publicly available with creative commons attribution license.

3 ETHICAL ASPECTS

3.1 PhenoMeNal Data managers and Data controller

In compliance with the Directive 95/46/EC and article 29 working group 8/2010 opinion (part of WP10 Ethics requirements), the consortium has appointed a Data controller responsible for monitoring data acquisition and management for the life of the project. The PhenoMeNal partners also nominated “data governance managers” from their group. The role of the data governance Managers is to obtain the relevant information on data used locally or distributed to other partners and to ensure, in collaboration with the Data Governance Manager, that it is ELSI compliant. The details of the role of data managers and data controller has been added as Annex 5.3

³⁸ <https://phenomenal-h2020.eu/jenkins/>

³⁹ <https://www.openaire.eu>



3.2 Data Availability process

A data availability process was established to define a reasonable process to make data available in PhenoMeNal. This includes the introduction of a registration process where the user is obliged to accept PhenoMeNal and ELSI “terms and conditions”. The process was an result of the guidance documents produced as project deliverables, ELSI workshops and meetings with the Ethics advisors. See deliverable 7.4 (process to extract maximum information from sensitive datasets with minimum compromise, in collaboration with BBMRI and BioMedBridges).⁴⁰

4 DATA SECURITY

4.1 Cloud tenancy

The PhenoMeNal infrastructure will only use compute resources available within the local tenancy. Access to local filesystems is controlled within the respective containers deployed in this tenancy, so direct remote filesystem access is not available. Primary research data uploaded directly to a deployed the PhenoMeNal infrastructure, ensure data reside locally within this tenancy only.

Public and private (embargoed and not ELSI restricted) data can be deposited and/or read from MetaboLights. Direct dataflow between MetaboLights and PhenoMeNal is controlled by, but not limited to, the MetaboLights-Labs-Uploader⁴¹ and MetaboLights-Labs-Downloader⁴² containers. These containers ensure secure transfer of data. Data uploaded directly to a running PhenoMeNal instance is under the specific user's control, so we are currently unable to limit uploads to the provided containers only.

4.2 Data preservation

All data submitted the public EMBL-EBI MetaboLights and EGA archives are stored securely across 3 geographical data centres. EMBL-EBI centrally funds parts of the operations of its data centers so data is securely preserved long term.

4.2.1 Data embargo and secure transfer

MetaboLights and EGA has existing access control mechanisms in place to prevent unauthorised access to embargoed and/or ELSI restricted data (EGA only).

PhenoMeNal have, together with the MetaboLights team, developed a secure high-performance data transfer mechanism for both embargoed and public data. This

⁴⁰ <http://phenomenal-h2020.eu/home/wp-content/uploads/2016/09/Deliverable7.4.pdf>

⁴¹ <https://github.com/phnmnl/container-mtbl-labs-uploader>

⁴² <https://github.com/phnmnl/container-scp-aspera>



mechanism is described in details in deliverable D9.2.2⁴³ (*PhenoMeNal-Data Virtual Machine image to enable sharing and dissemination of standardised and processed omics data to participating online repositories, like MetaboLights*)

For ELSI restricted data PhenoMeNal will use the EGA archive, and the encrypted data transfer mechanisms⁴⁴ already established. Data is encrypted by the submitter and further re-encrypted as part of the EGA archive process.

⁴³ <http://phenomenal-h2020.eu/home/wp-content/uploads/2016/09/D9.2.2PhenoMeNal-DataVirtualmachine.pdf>

⁴⁴ <https://www.ebi.ac.uk/ega/about/security>



5 ANNEXES

5.1 Renewed permission to use MESA dataset

-----Original Message-----

From: W. C. Johnson [<mailto:wcraigj@uw.edu>]

Sent: 08 February 2017 16:01

To: Ebbels, Timothy M D <t.ebbels@imperial.ac.uk>; David Vu <voodoo@uw.edu>

Cc: Kayleen Williams <kmfw@uw.edu>; Glen, Robert C <r.glen@imperial.ac.uk>; Pearce, Jake T M

<jake.pearce@imperial.ac.uk>; 'Gregory L. Burke' <gburke@wakehealth.edu>; 'Russell P Tracy'

<russell.tracy@med.uvm.edu>; 'Jerome I. Rotter' <jrotter@labiomed.org>; 'David Herrington' <dherring@wakehealth.edu>;

'Philip Greenland' <p.greenland@northwestern.edu>

Subject: RE: MESA use case for PhenoMeNal?

Hi Tim,

My apologies for the delayed response. Continued use of the MESA data is fine while we work on renewal of the MESA DMDA with Imperial College. David Vu from our offices here will be in touch with the template DMDA form and instructions.

David - Once you have sent off the documents to Dr. Ebbels, I would appreciate a review of the status of DMDAs we have on file with the other COMBI-Bio Institutions (just to see whether/what additional DMDA renewals might be needed. Finally, we will need to reopen discussion with NCBI concerning upload/posting of the COMBI-Bio project metabolomics datasets on dbGaP.

Thank you,

Craig

Craig Johnson
Project Director, MESA CC
Associate Director of Research, CHSCC
University of Washington
Collaborative Health Studies Coordinating Center Building 29, Suite 310
6200 NE 74th Street
Seattle, WA 98115-8160

Phone: 206-897-1911
Office: 206-685-7123
Cell: 425-418-2822
Fax: 206-616-4075



5.2 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

5.3 Guidance for Data Governance Managers

In the award of the grant from the EU, Phenomenal was asked to consider carefully any ELSI (Ethical, Legal and Social Implications) for the use of human data. This includes data which is anonymised, is open, restricted, private etc. ELSI should be considered for all cases of data use.

Phenomenal is an infrastructure project to process data. It is not for storing data. During the development phase we will use data to evaluate the performance of Phenomenal and we will need to consider the ELSI implications of this. The overall project has a Data Controller (Glen) and each site has a Data Manager. The role of the Data Governance Managers is to obtain the relevant information on data used locally or distributed to other partners and to ensure, in collaboration with the Data Governance Manager, that it is ELSI compliant.

In Deliverables (7.1, 7.2, 7.3, 7.4, 7.5) we have considered ELSI requirements in Phenomenal. As your local Data Manager you should read and understand these Deliverables as this will help to securely manage the data that we will use to test Phenomenal infrastructure. You are responsible for the use (and any abuse) of data you provide or use in Phenomenal. If you intend to use data, you should be aware of the terms of use – for example, you may use open data from Metabolites, which comes under the EBI terms of use, or other data may have restrictions on how it can be used or distributed. If you provide data e.g. for testing proposes by another group in Phenomenal, you should follow the guidelines in the Deliverables (flowcharts, terms of use of Phenomenal, data submission forms) to make the data available under the ELSI constraints and consent given for the use of this data at your institution. We will hold a registry of the data used in Phenomenal and also the consent that goes with the data. The consent may be a number of documents including ethical approval, patient consent, terms of use etc. Before using or making data available, we should be satisfied that what we are doing is compliant.

The following information will be needed:

1. Please supply a description of the data
 - a. Who is the main contact for the data (should be a Data Governance Manager)
 - b. What is the data
 - c. Source of the data



- d. Is it open or restricted
 - e. Does it contain patient metadata
 - f. Can any individual be identified from the data
 - g. Where will the data be used (local institution or distribution list)
2. Please supply the consent, NMOU's, DACS's and other documentation that goes with the data and confirm that the data is being used in accordance with these. This will be stored along with the data description.

List of Data Governance Managers in PhenoMeNal

PARTNER	DATA MANAGER
EMBL-EBI	Kenneth Haug
ICL	Robert Glen
IPB	Daniel Schober
UB	Pedro
UoB	Karen Atkins
CIRMMP	Leonardo Tenori
UL	Michael van Vliet
UOXF	Philippe Rocca-Serra
SIB	Sven Bergmann
UU	Kim Kultima
BBMRI	Petr Holub
CEA	Etienne Thévenot
INRA	Fabien Jourdan



CRS4	Marco Enrico Piras
------	--------------------



6 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