Deliverable 1.5.3

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	27th February 2018
Version	3.0
Delivery Date	M30
Work Package leader	EMBL-EBI
Contributing Partners	EMBL-EBI
Authors	Namrata Kale, Ken Haug, Philippe Rocca-Serra, Kim Kultima, Etienne, 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 **<u>Changes from D1.5.1 - 15.02.2017:</u>** ● Removed references to Mosler as a preferred data repository ● URLs as footnotes only ● Added FAIR data clarification ● Added metabolomics study size information ● Added data security ● Removed references to any use of the CoLaus due to consent to use data ● Information on the data produced/generated during the project **<u>Changes from D1.5.2 - 20.07.2017:</u>** ● Added paragraphs for non-scientific data, a.k.a. “Other Data” **<u>Changes from revised D1.5.2 - 18.08.2017:</u>** ● GDPR preparation

1 EXECUTIVE SUMMARY

The PhenoMeNal project has developed and deployed 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. PhenoMeNal is addressing 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 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 provides 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 is 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 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 a revised deliverables D1.5.3 incorporating General Data Protection Regulation (GDPR).

2 DATA SUMMARY

PhenoMeNal is an e-infrastructure for managing, preserving and computing with biomedical phenotyping in combination with genomic data that phenome centres and biomedical laboratories generate from human research subjects. Thus, within the PhenoMeNal project, the research data is deposited by the users, a.k.a. "data providers", of the infrastructure, prior to externally generating and producing said data. The project facilitates the storage of relevant data in secure public environments using data standards and procedures developed by **CO**ordination Of **S**tandards In Metab**O**lomic**S** (COSMOS[1]), the Metabolomics Standards Initiative (MSI[2]) and European Translational Information & Knowledge Management Services (eTRIKS[3]). Metagenomic, genomic and metabolomics data and protocols will, where appropriate, be deposited with the European Bioinformatics Institute (EMBL-EBI, UK).

EMBL-EBI MetaboLights[4], is the integrated data repository for PhenoMeNal. The average study collection size in MetaboLights is about 25Gb[5], however the median size is around 1.1Gb per study. Due to the homogeneity of metabolomics data, MetaboLights accepts all types of metabolomics data from the worldwide metabolomics community. All public data in MetaboLights is freely available for any individual and for any purpose, as per EMBL-EBI guidelines[6]. In PhenoMeNal, useable data types are very much dependant on the capabilities of the chosen workflow and/or tools therein. We have endeavoured to describe these capabilities in

[1] http://www.cosmos-fp7.eu/
[2] http://www.metabolomics-msi.org/
[3] https://www.etriks.org/
[4] https://www.ebi.ac.uk/metabolights/
[5] https://www.ebi.ac.uk/metabolights/statistics
[6] https://www.ebi.ac.uk/about/terms-of-use

the technical description available through the Service Catalogue[7] 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 initial use of restricted data is for development and testing of software and computational tools, but no novel biological analysis will be performed or published.

We initially had additional use cases involving datasets from Switzerland (CoLaus) and Sweden (Uppsala Fibromyalgia) but data consent was not given.

The MESA dataset

The Multi-Ethnic Study of Atherosclerosis, MESA[8], 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[9]), 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 do not impose any additional metadata requirements for privacy restricted datasets. These datasets are governed by their existing ELSI requirements, which has 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

[7] http://portal.phenomenal-h2020.eu/app-library
[8] http://www.mesa-nhlbi.org/
[9] http://www.combi-bio.eu/

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)[10] at the EMBL-EBI. The data will not be public and will be accessible, through normal EGA data access procedures (Data Access Committee), only to the members of the PhenoMeNal consortium for testing purposes. PhenoMeNal and EGA will not impose additional metadata requirements for these datasets.

Uppsala Fibromyalgia study

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

Standards and metadata

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

Data sharing

The target audience for these restricted datasets are the PhenoMeNal researchers within Akademiska Sjukhuset (Uppsala) for testing and none of this data will be made publicly available. The use of the data is 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. PhenoMeNal and EGA will not impose additional metadata requirements for these datasets

[10] https://www.ebi.ac.uk/ega/

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 is also 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 lead initiatives like COSMOS FP7 and the MetaboLights project (BBSRC). MetaboLights require all datasets to be MSI compliant and annotated using the ISA-Tab format[11]. The ISA-Tab format, endorsed by several data repositories (Metabolights, GigaDB[12], Dryad[13]) and publishers (Nature Publishing Group, Oxford University Press) is now supported by an open API[14] and visualisation/search tool[15]. The ISA metadata format, and the MSI approved open access raw data formats (mzML and nmrML) used in 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 (templates) and validation rules, which will result in regularised and augmented semantic mark up. This latter step will also increase the number of entities made identifiable and therefore web 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.

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 analysed by liquid chromatography coupled to high-resolution mass spectrometry[16].

Within PhenoMeNal, the objectives are:

- To reproduce the publicly available reference workflow (W4M00001_Sacurine-statistics[17]; Workflow4Metabolomics[18] e-infrastructure) on the PhenoMeNal cloud environment,

[11] http://isa-tools.org/format/specification/
[12] http://gigadb.org
[13] http://datadryad.org
[14] https://github.com/ISA-tools/isa-api
[15] http://isa-explorer.org
[16] Thévenot et al, 2015, http://pubs.acs.org/doi/full/10.1021/acs.jproteome.5b00354
[17] http://dx.doi.org/10.15454/1.4811121736910142E12
[18] http://workflow4metabolomics.org/

- 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. The International Metabolomics Society[19] are about to publish a new, up-to-date, standard for defining and reporting evidence levels.

Data sharing

Raw data are publicly available in the Workflow4Metabolomics e-infrastructure. Within the PhenoMeNal project, the full data set is also made available in the EMBL-EBI MetaboLights repository as study MTBLS404[20].

Ethical and legal compliance

This is an open publicly available dataset, which is 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 long term (MetaboHUB[21]: National Infrastructure for Metabolomics and Fluxomics; and IFB[22]: French Institute of Bioinformatics - Elixir node). In addition, in accordance with EMBL policy, the operation and running of the strategic EBI MetaboLights archive is centrally funded and will be maintained without the need for short or medium-term funding.

Data from fluxomics analysis

Determination of metabolic flux distributions is fundamental in order to have a complete characterisation of metabolic phenotypes. During these analysis, cells are incubated with isotope-enriched substrates to decipher their biochemical processing through the main metabolic pathways. This type of analysis 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[23]. 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

[19] http://metabolomicssociety.org/
[20] http://www.ebi.ac.uk/metabolights/MTBLS404
[21] http://www.metabohub.fr/home.html
[22] http://www.france-bioinformatique.fr/en
[23] 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#

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[24]. 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 is 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 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 is publicly available, where applicable. 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.

2.1.2 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 technique. The types of data include experimental NMR, LC-MS, GC-MS, Imaging and chromatographic data.

[24] http://www.metabolomicsworkbench.org

Standards and metadata

The MetaboLights submission pipeline is utilising the ISA software suite[25]. 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 January 2018, MetaboLights holds over 185 human datasets, of which about 55% is in the public domain.

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 VRE[26] 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. 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 repositories 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 situ until the user actively removes it or exports it to another 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 within or outside the VRE, although it does not prohibit the user's ability to do so. User

[25] http://isa-tools.org/
[26] http://portal.phenomenal-h2020.eu/home

controlled personal VREs where the owner removes their own instance, will naturally not retain any data.

2.1.3.1 Public Test VRE

PhenoMeNal offers 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. User registration terms and conditions, as well as the relevant online forms to do so, are being rewritten/redeveloped to comply with GDPR when this comes into force in May 2018.

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 Galaxy[27] account, on the PhenoMeNal infrastructure, or with a supported public cloud supplier, are able to control the data therein. As previously mentioned; T&Cs, as well as the relevant online forms, are currently changing to comply with the upcoming GDPR. 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, has different requirements for what type of

[27] http://public.phenomenal-h2020.eu

data files can be processed. Each tool and compatible file formats is 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 (online server environments) within the PhenoMeNal infrastructure will at some point have to be time limited to enable fair sharing of available resources, this is not a current constraint. This is obviously not the case where the user has deployed the VRE to a commercial cloud provider, like Amazon or Google. 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.

Standards and Metadata

PhenoMeNal is not imposing any additional requirements related to metadata for datasets. These datasets are governed by their existing privacy and ethical requirements, and metadata requirements has been defined during the design of each study.

Data sharing

PhenoMeNal does not, and will not, offer direct data sharing from within any VRE, however some tools facilitate mechanisms to publish the uploaded and/or derived data to the participating data repositories. Guidance is provided for data depositions into the MetaboLights repository, which is currently the only deposition service linked to the PhenoMeNal pipeline. MetaboLights, and the data therein, is governed by EMBL-EBI 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 is after consent only and anonymised data should be used whenever possible. This is governed by PhenoMeNal terms of use[28]. See Annex 5.2

[28] http://phenomenal-h2020.eu/home/wp-content/uploads/2016/09/Phenomenal-Terms-of-Use-version-11.pdf

Archiving and preservation

This is considered beyond this scope as this is governed by the respective repositories existing 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 a user controlled VRE where explicit permission 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. As this type of VRE is entirely controlled by the respective user, no further restrictions or controls can be enforced by the PhenoMeNal consortia.

Standards and Metadata

PhenoMeNal does 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, 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. Where this data relates to a specific

user, i.e. a user has to log into some part the VRE, we are developing policies and making appropriate software changes in accordance with the upcoming GDPR.

2.2.1.1 User accounts

2.2.1.1.1 Public accounts

PhenoMeNal has co-developed a new single sign-on (SSO) system[29] 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 id[30], 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, 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, i.e. 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.

This is an area where PhenoMeNal are currently focusing significant efforts to comply with GDPR. Elixir and EMBL are making required changes to the public SSO solution to support the upcoming regulations. Additionally, VRE local Galaxy and Jupyter user accounts, and the PhenoMeNal internal user management system will require policy and software changes.

2.2.1.1.2 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.
When deploying and using a PhenoMeNal VRE in external, to the project, cloud providers, any *potential additional* GDPR implications must be considered. It is assumed that local governments, and commercial companies operating within the

[29] https://www.elixir-europe.org/services/compute/aai
[30] username@elixir-europe.org

same geographical boundaries, can apply additional restrictions in addition to the current "GDPR baseline". These restrictions may dictate additional changes to terms & conditions and any software (user accounts, infrastructure, workflow, tool or process) available within, or underpinning, a PhenoMeNal VRE. The PhenoMeNal consortia working towards compliance with the general EU GDPR, but not beyond this scope.

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. GDPR will naturally apply and required changes are being investigated.

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. GDPR applies and operational procedures will have to change accordingly.

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 (registered users only) 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 selected 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 Zenodo[31], in addition to the project website[32], as agreed with the commission. The consortia is currently investigating how GDPR will affect the use of online sharing services like Google Drive.

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-EBIs private Google Suite ensuring only invited project members can join. In addition to the extra security for 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 project[33] for the majority of direct, one-to-one or group, communication. This greatly reduces the volume of emails in the project and

[31] https://zenodo.org/communities/phenomenalh2020/
[32] http://phenomenal-h2020.eu/home/about/objectives/
[33] https://phenomenal.slack.com

ensures effective discussions. The PhenoMeNal consortia is currently investigating if GDPR will affect how we use Slack.

2.2.1.5 Project Management

PhenoMeNal is using Pivotal Tracker[34] 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. The PhenoMeNal consortia is currently investigating if GDPR will affect how we use Pivotal Tracker.

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 GitHub[35] 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. The PhenoMeNal consortia is currently investigating if GDPR will affect how we use 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, or 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. GDPR obviously applies and terms & conditions will have to change according to the upcoming regulations.

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.

2.2.2.4 Continuous integration

We use a continuous integration (CI) system[36] 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,

[34] http://phenomenal-h2020.eu/home/about/objectives/project-management-tool/
[35] https://github.com/phnmnl
[36] https://phenomenal-h2020.eu/jenkins/

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. User accounts associated with the CI system falls under GDPR and appropriate changes to PhenoMeNal terms and conditions applies.

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 OpenAire[37]. Video demonstrations available on YouTube are publicly available with creative commons attribution licence.

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

3.2 Data Availability process

A data availability process was established to define a reasonable process to make data available in PhenoMeNal. The process was an outcome of the guidance documents produced as deliverables, ELSI workshops and meetings with the Ethics advisors.

4 DATA SECURITY

4.1 Cloud tenancy

The PhenoMeNal infrastructure will only use compute resources available within the local tenancy in which a PhenoMeNal VRE is deployed. Access to any local

[37] https://www.openaire.eu

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 version of the PhenoMeNal infrastructure, ensures 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 data-flow between MetaboLights and PhenoMeNal is controlled by, but not limited to, the MetaboLights-Labs-Uploader[38] and MetaboLights-Labs-Downloader[39] 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 not limiting 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 centres so data is securely preserved long term.

4.3 Data embargo and secure transfer

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

PhenoMeNal have, together with additional members of the MetaboLights team, developed a secure high-performance data transfer mechanism for both embargoed (initial access has to be specifically facilitated) and public data. This mechanism is described in details in deliverable D9.2.2[40] (*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 mechanisms[41] already established. Data is encrypted by the submitter and further re-encrypted as part of the EGA archive process.

[38] https://github.com/phnmnl/container-mtbl-labs-uploader
[39] https://github.com/phnmnl/container-scp-aspera
[40] http://phenomenal-h2020.eu/home/wp-content/uploads/2016/09/D9.2.2PhenoMeNal-DataVirtualmachine.pdf
[41] 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 reflect 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

6.1 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.

6.1.1

6.1.2 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

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