
# Deliverable 9.2.2

| | |
|---|---|
| **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** | 9 |
| **Work Package Title** | WP9 Tools, Workflows, Audit and Data Management |
| **Deliverable Title** | D9.2.2 *PhenoMeNal*-Data Virtual Machine image to enable sharing and dissemination of standardised and processed omics data to participating online repositories, like MetaboLights |
| **Delivery Date** | M14 |
| **Work Package leader** | IPB |
| **Contributing Partners** | EMBL-EBI, IPB, UOXF, ICL, CEA, INRA, UB, CRS4 |
| **Authors** | Ken Haug, Pablo Moreno, Venkata Chandrasekhar, Reza Salek, Steffen Neumann, David Johnson, Kristian Peters, Daniel Schober, Jianliang Gao, Alejandra Gonzalez-Beltran, Luca Pireddu, Noureddin Sadawi |

**Abstract:** In this deliverable we report the handling of primary research data files (raw data) in the PhenoMeNal Virtual Research Environment (VRE) and the open MetaboLights repository – one of the two major Metabolomics data repositories.

## TABLE OF CONTENTS

# 1. EXECUTIVE SUMMARY

In this deliverable (*D9.2.2 PhenoMeNal-Data Virtual Machine image to enable sharing and dissemination of standardised and processed omics data to participating online repositories)*, we report on the handling of primary research data files (raw data) in the PhenoMeNal Virtual Research Environment (VRE) and the open MetaboLights repository[1] – one of the two major Metabolomics data repositories. We provide easy-to-use mechanisms for fast and secure data transfer between the VRE and MetaboLights.

# 2. WORK TOWARDS PROJECT OBJECTIVES

A summary of work towards the project objectives:

**Objective 9.1**: "Specify and integrate software pipelines and tools utilised in the PhenoMeNal e-Infrastructure into VMIs, adhering to data standards developed in WP8 and supporting the interoperability and federation middleware developed in WP5. Most tools will be already available (see table 1.1) and we will develop new applications to complete 'missing links' in pipelines. Although two explicit releases for VMIs are listed as deliverables below, we will use public repositories and continuous integration to always provide development snapshots of the infrastructure VMIs."

- Specified and integrated two pipelines: one for data imported from MetaboLights (including the data transfer mechanisms), and a second pipeline for data to be submitted to MetaboLights.
- Tools packaged in containers (VMIs equivalents): mzml2isa, ISA slicer, MetaboLights Downloader, MetaboLights Labs Uploader. Other tools have already been reported in previous deliverables.
- Most of the tools, where applicable, support ISA-Tab metadata files.
- All the packaged tools are available on the PhenoMeNal public docker registry.

**Objective 9.2**: "Develop methods to scale-up software pipelines for high-throughput analysis, supporting distributed execution on e.g. local clusters, private clouds, federated clouds, or GRIDs."

[1] http://www.ebi.ac.uk/metabolights/

- All of the software packaged in containers has been tested to run on scalable infrastructure (Kubernetes container orchestrator on EMBL-EBI EMBASSY OpenStack installation).
- We have taken steps to make importing large datasets into PhenoMeNal feasible through the use of advanced data transfer clients and methods that avoid transferring unneeded data by slicing datasets and retrieving only the required portions.

# 3. DETAILED REPORT ON THE DELIVERABLE

Here we report on the handling of primary research data files (raw data) in the PhenoMeNal Virtual Research Environment (VRE) and the open MetaboLights repository as well as the data communication between these two. We supply easy-to-use mechanisms for fast and secure data transfer between the VRE and MetaboLights, as an example of one of the two major Metabolomics data repositories. Experimental metadata is organised in the ISA-Tab[2] format and captures information on the level of the **I**nvestigation, **S**tudy and **A**ssays. Privacy considerations govern the decision which data can be analysed in local, protected or public repositories.

## 3.1. Introduction

It is essential to enable simple mechanisms to transfer data from the PhenoMeNal VRE and data repositories such as the MetaboLights repository:

1. From PhenoMeNal to MetaboLights: submitting data analysed within the PhenoMeNal VRE into MetaboLights.
2. From MetaboLights to PhenoMeNal: re-analysis of data available in MetaboLights in PhenoMeNal.

MetaboLights has a new feature known as “MetaboLights Labs[3]”, which in the context of PhenoMeNal can be viewed as a staging and pre-processing area for primary data. By staging in this context we mean the addition of data to a preliminary storage area of MetaboLights, where the user can work with the uploaded data until he decides to promote datasets as a new study for proper submission. The user ultimately controls this data, so if the end goal or later decision is not to publish as a new study, the user is

[2] http://isa-tools.org
[3] www.ebi.ac.uk/metabolights/labs

free to delete it from the staging area, which is not possible for the permanent data published on MetaboLights. MetaboLights Labs is mainly intended to be used for raw data upload before data is analysed and/or processed into a complete study. It uses the concepts of a user "Workspace" consisting of one or more "Projects", each offering a feature to progress a dataset into a complete MetaboLights study. This staging mechanism is non-committal for the user in terms of the more extensive annotation requirements in MetaboLights, but already allows to prepare and manage the data as a study. Should the user choose to publish data as a fully annotated MetaboLights study, the process is handled through the normal submission and curation procedures in MetaboLights.

To analyse data within the PhenoMeNal platform, the user can choose to work through the Galaxy framework directly within the VRE. In this case, data can be made available for analysis in the following two ways:

1. Directly available to PhenoMeNal VRE Galaxy framework. Raw data (aka. primary data) can be uploaded directly from the user's local file system/data ingestion point.
2. Indirectly available to PhenoMeNal VRE Galaxy framework. Raw data can be uploaded to MetaboLights, from where it can be subsequently downloaded to PhenoMeNal VRE Galaxy installation. In order to upload data to MetaboLights, raw data can be ISA-Tab annotated studies or raw files only as part of a MetaboLights Labs project.

Metabolomics datasets can be rather large, and thus time-consuming and resource-intensive (in terms of network and storage capacity) to handle. If only a subset of the dataset is required for an analysis, depending on age, sex, treatment or other factors, data slicing allows to transfer only the relevant portion of the data. The required metadata is organised in the ISA-Tab format and captures information on the level of the **I**nvestigation, **S**tudy and **A**ssays. PhenoMeNal VRE users can choose datasets and associated raw files based on experimental factors or characteristics of materials and data using the "mtblisa" tool described in this deliverable.

### 3.2. Access restriction considerations

For patient data, sharing and reuse must be governed by ethical, legal and environmental implications (ELSI). Where restrictions exist in terms of ELSI, data submission to a public archive or VRE is not allowed. The European Genome-Phenome Archive (EGA)[4] is the preferred data archive at EMBL-EBI for privacy restricted studies.

[4] https://www.ebi.ac.uk/ega/home

If one intends to use the PhenoMeNal VRE for restricted data that cannot be uploaded into a repository at the EBI, the VRE can be installed and run within a private network. In that 'bring-the-compute-to-the-data' scenario, local users will upload data directly to their local PhenoMeNal VRE. These installations are supported and documented in WP5 (Operations and Maintenance of PhenoMeNal GRID/CLOUD).

Additionally, access restrictions can also result from a temporary embargo – e.g., until a predetermined publication date has been reached. This scenario is well handled by the MetaboLights data management capabilities.

In the context of this document, we focus on the aspect of data management and transfer in the context of the VRE and the MetaboLights repository.

Datasets exist in MetaboLights in two main forms: public and private. Public data exists in the form of complete studies that, when running a workflow in the VRE, are already available online in MetaboLights without any restrictions. Private data are only available to researchers that have credentials to access it. In MetaboLights, credentials are given on a per-project/user basis. Private data can be date embargoed, typically for pre-publication purposes.

Once data have been downloaded from MetaboLights to the PhenoMeNal VRE, they are only accessible to the local Galaxy user who triggered the download, unless the user actively shares the dataset with other users in the same Galaxy VRE instance. We do not seek to facilitate direct sharing with users external to the PhenoMeNal VRE Galaxy instance, and rely on the data transfer to MetaboLights as dedicated repository if sharing is intended.

The administrator who has deployed a local VRE must limit the set of users with administrator credentials to staff that strictly needs access, to prevent unauthorised access to data in the underlying VMs through Linux shell commands. This is mentioned in the installation instructions[5], but naturally goes beyond the control of the PhenoMeNal consortium and falls under the regular user management practices of each private VRE installation. If no administrative access is given to untrusted users, we have no technical reason to think that the data would be at risk of unauthorized access.

[5] https://github.com/phnmnl/phenomenal-h2020/wiki/QuickStart-Installation-for-Local-PhenoMeNal-Workflow

### 3.3. Data exchange architecture between MetaboLights and PhenoMeNal VRE

Here we describe in more technical detail how the bi-directional dataflow will be facilitated between MetaboLights and the PhenoMeNal VRE, through our Galaxy workflows. Figure 1 shows the conceptual dataflow.

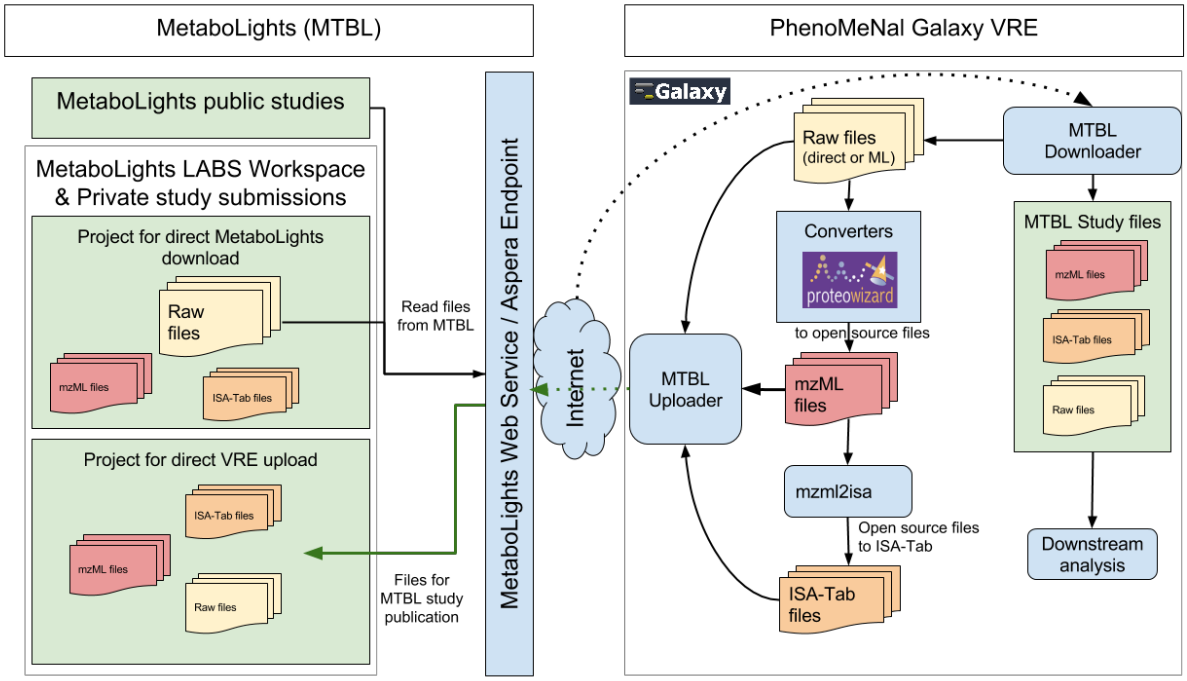


**Figure 1:** Data flow for Mass Spectrometry (MS) workflow use cases. Both raw files and ISA-Tab archives can be uploaded either directly to the PhenoMeNal VRE through the Galaxy interface (right), or transferred from MetaboLights and MetaboLights LABS (on the left) via the MTBL Downloader. MS raw files can be converted into open format mzML files using ProteoWizard[6]. The "mzml2isa[7]" tool supports conversion from open formats MS (mzML), MS Imaging (imzML) and NMR (nmrML) to ISA-Tab files. After analysis, the automatic study submission transfers data to MetaboLights LABS (green arrow).

[6] http://proteowizard.sourceforge.net/
[7] https://github.com/althonos/mzml2isa

### 3.3.1. MetaboLights HTTP, FTP and Aspera incoming data flow

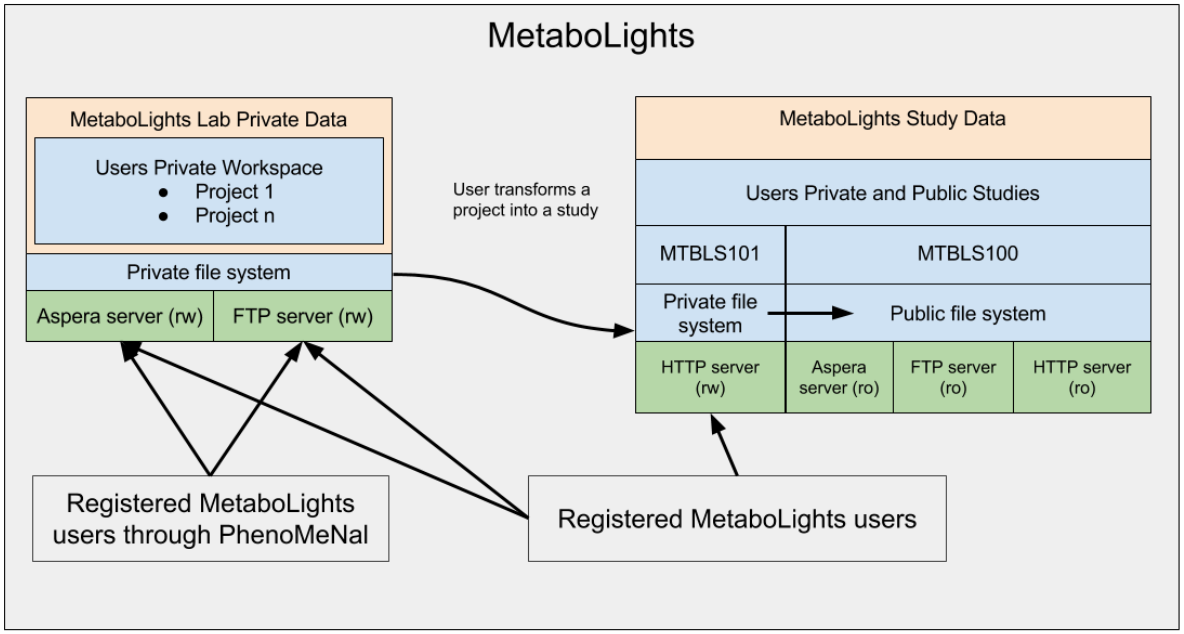


**Figure 2:** Overview of MetaboLights' server architecture pertaining to data coming in from registered users. Data transfer endpoints (green) can be either read-write (rw) to registered and authorized users, or read-only (ro) to all users after publication.

### 3.3.2. Support of highly efficient Aspera data transfer protocol

To facilitate the reliable and fast transfer of large datasets, we have chosen to implement a solution using Aspera transfer clients. The Aspera software uses its custom Fast, Adaptive and Secure Protocol (FASP) on top of the omnipresent IP protocol, replacing TCP and upper layers of the network stack (e.g., HTTP, FTP). Compared with conventional data transfer mechanisms, the system provides faster and more reliable data transfer over typical Internet links where communication latency is higher than in a typical local LAN. As an example, the data transfer for a 4.5 GB dataset (MTBLS20[8]) with Aspera takes less than half the time over a relatively fast connection[9], compared to FTP. It should be noted that over a slower connection, the choice of transfer protocols has less influence. See Annex 1 for an example. Cloud infrastructures can be expected to have good internal and external bandwidth available.

[8] http://www.ebi.ac.uk/metabolights/MTBLS20
[9] 850Mbps

### 3.4. Data transfer from MetaboLights to the PhenoMeNal VRE

In this scenario raw data is transferred in the form of a MetaboLights study from MetaboLights to the PhenoMeNal VRE. As mentioned earlier, data in MetaboLights is either public or private. We first describe access to both types, and then describe the PhenoMeNal solutions to integrate these into the VRE.

#### 3.4.1. MetaboLights Public data download

A common use case for the PhenoMeNal VRE is secondary data analysis – e.g., to re-analyse publicly available metabolomics datasets. Public data submitted to MetaboLights is, per EMBL-EBI licensing, open to anyone for all purposes. MetaboLights facilitates data transfer using a few different protocols, all freely available over the Internet. Typical protocols/methods of transfer are FTP, HTTP and Aspera[10]. Clients for these protocols include popular tools such as wget and curl.

Understanding that we do not aim to expose the end user to this level of technicality, for the sake of illustration of what we are wrapping, these are command line examples for including MetaboLights data from the public MTBLS20[11] study.

FTP:

```
$ ftp -in ftp://ftp.ebi.ac.uk/pub/databases/metabolights/studies/public/MTBLS20 << EOF
user anonymous password
binary
mget *
EOF
```

wget (using FTP or HTTP server):

```
$ wget ftp://ftp.ebi.ac.uk/pub/databases/metabolights/studies/public/MTBLS20/*
or
$ wget http://www.ebi.ac.uk/metabolights/MTBLS20/files/MTBLS20

For ISA-Tab metadata only you can use:
$ wget http://www.ebi.ac.uk/metabolights/MTBLS20/files/metadata
```

Aspera:

[10] http://asperasoft.com/
[11] http://www.ebi.ac.uk/metabolights/MTBLS20

```
$ export ASPERA_SCP_PASS=<password>
$ ascp -QT -l 1g fasp-ml@fasp.ebi.ac.uk:/studies/public/MTBLS20 .
```

Command line Aspera clients can be freely downloaded from the Aspera website[12].

### 3.4.2. MetaboLights Private data download

Another use case is the analysis of anonymised human datasets that are either under pre-publication embargo or intended for only a smaller audience with direct access. Studies are naturally not available to public access so there are fewer options for data sharing.

Again, the following command line examples for including private data from MetaboLights show the functionality that is available to the end-user in the Galaxy tool. We use the Aspera client[13] to read data from MetaboLights Labs or private Aspera (and FTP) server:

```
$ export ASPERA_SCP_PASS=<password>
$ ascp -QT -l 1g mtblight@ah01.ebi.ac.uk:<folder-obfuscation-location> .
```

MetaboLights Labs is considered a staging and analysis area for new studies in MetaboLights. To access an already existing private study folder, the command shown below is executed by the Galaxy tool.

HTTP (using wget) from MetaboLights web service for private study:

```
$ wget http://www.ebi.ac.uk/metabolights/MTBLS353/files/MTBLS353?token=<obfuscation-code>

For ISA-Tab metadata only you can use:
$ wget http://www.ebi.ac.uk/metabolights/MTBLS353/files/metadata?token=<obfuscation-code>
```

### 3.4.3. PhenoMeNal Container and Galaxy wrapper for downloader

We encapsulate the required command line utility to download data from MetaboLights to the Galaxy VRE in a Galaxy wrapper named mtbls-downloader – available in the PhenoMeNal public instance (http://public.phenomenal-h2020.eu, currently in the Tool's

[12] http://downloads.asperasoft.com/en/downloads/2

[13] http://download.asperasoft.com/download/sw/ascp-client/3.5.4/ascp-install-3.5.4.102989-linux-64.sh

section “Transfer”). Figure 3 shows mtbls-downloader’s user interface embedded in the PhenoMeNal Galaxy VRE, as well as the section in Galaxy’s tool library where the uploader can be found (on the left). Figure 4 shows the result of the execution of the tool when downloading data.

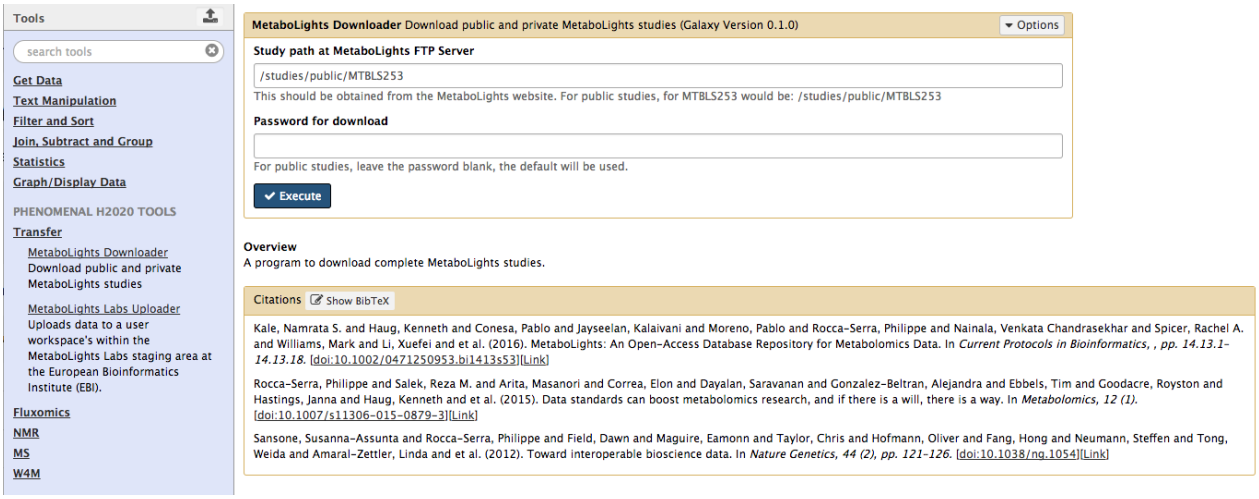


**Figure 3:** The MetaboLights Downloader Galaxy module. The tool allows users to retrieve both public and private studies from MetaboLights. For public studies, a default password is used and indications are given to the user on how to fill in the adequate path for a study. Currently, the user can find the MetaboLights Downloader in the Transfer section of the PhenoMeNal H2020 Tools, on the left side of the screen in the purple panel.


**Figure 4:** The result of running the MetaboLights Downloader module on a Galaxy VRE. Produced files are available on the right side of the screen, under “History”. The tool can produce both Zipped and individual versions of MAF (Metabolite Annotation File) and ISA files, for different downstream usages in Galaxy, as well as a TAR archive file containing all the data files obtained from MetaboLights for that study. By default, these files are only visible to the user that downloaded them; they need to be actively shared to become visible to other users within Galaxy.

The MetaboLights Downloader module relies on a container written for this purpose which packages the Aspera download mentioned previously. The container build file is available at github.com/phnmnl/container-scp-aspera. The container image can be obtained from our PhenoMeNal docker registry through:

```
docker pull docker-registry.phenomenal-h2020.eu/phnmnl/scp-aspera
```

The Aspera client can be used from the docker container (on a machine running docker) through:

```
docker run --env ASPERA_SCP_PASS=<password> -w=$PWD -v /home/your_user:/home/your_user \
docker-registry.phenomenal-h2020.eu/phnmnl/scp-aspera -QT -l 1g \ fasp-
ml@fasp.ebi.ac.uk:/studies/public/MTBLS<number>
```

The MetaboLights Downloader Galaxy module has been tested on the PhenoMeNal public Galaxy VRE, running on top of Kubernetes.

### 3.5. Data transfer from the PhenoMeNal VRE to MetaboLights Labs

In this scenario raw data is uploaded directly to the PhenoMeNal VRE from the user's local filesystem or data ingestion point through the existing Galaxy upload facilities. We use ProteoWizard msconvert and mzml2isa, available as tools in our workflow, to convert from vendor-specific to open file formats. The converters have been described as part of D9.2.1. These open formats – typically mzML – are then automatically parsed and mapped into annotated ISA-Tab documents. These ISA-Tab documents are ready to be submitted to the MetaboLights Labs staging area as a private study. After the analysis, PhenoMeNal offers the user the ability to transfer all relevant files, raw and processed, into a MetaboLights Labs project. This is outlined in Figure 1. The study will then be further curated within the context of MetaboLights.

#### 3.5.1. PhenoMeNal to MetaboLights Python Uploader

To execute the data transfer from the PhenoMeNal VRE Galaxy framework to MetaboLights LABS, we have created a command line utility that encapsulates the more complex underlying commands described in previous sections. This tool is implemented as a wrapper[14] written in Python[15], which is called from the Galaxy user interface to enable quick and secure transfer of data from a user VRE Galaxy workspace area into a MetaboLights Labs project.

[14] https://github.com/EBI-Metabolights/MetaboLightsLabs-PythonCLI
[15] https://www.python.org/

Transferred data is under user control and governed by existing EMBL-EBI MetaboLights guidelines. The user can at this point choose to publish this data as a new study, using existing MetaboLights functionality.

The following illustrates the technical details wrapped as a Galaxy tool for uploading data to MetaboLights Labs:

```
$ upload_to_labs.py -t api_key_string -i pathToFile1 … pathToFileN [ -p labs_project_id -n ]

Parameters:
       -t  : User provided MetaboLights API key. This key is available to all registered
       users in MetaboLights under the user's personal account page
       -i :  A space separated list of files and/or folders that should be transferred to
       MetaboLights Labs
       [-p : The id of the users unique MetaboLights Labs project. To create a new project
       you can specify the '-n' parameter and the project will be created under the given
       name. If no name has been supplied, a default project will be created.]
       [-n : Create new project if the name supplied in '-p' does not exist. Can only be
       used in conjunction with the '-p' parameter.]
```

- Parameters '-t' and '-i' are mandatory. If no '-p' parameter is supplied a brand new workspace (if none exist) and a project will be created in MetaboLights Labs for the user.
- The list for the '-i' parameter should contain all the files and/or folders that are relevant for a future MetaboLights study: raw files, open source mzML files and ISA-Tab files.
- Requirement: An approved MetaboLights user account

### 3.5.2. PhenoMeNal Container and Galaxy wrapper for uploader

To call the command line utility described above from Galaxy in the VRE, we provide the Galaxy wrapper mtbls-labs-uploader, available on GitHub (https://goo.gl/E1zsoF), as part of and pre-installed in the container-galaxy-k8s-runtime container. The tool is thus also available in the PhenoMeNal public instance (http://public.phenomenal-h2020.eu, in the tool section "Transfer"). Figure 5 shows the user's view in Galaxy, as well as the section of Galaxy tools where the uploader can be found.

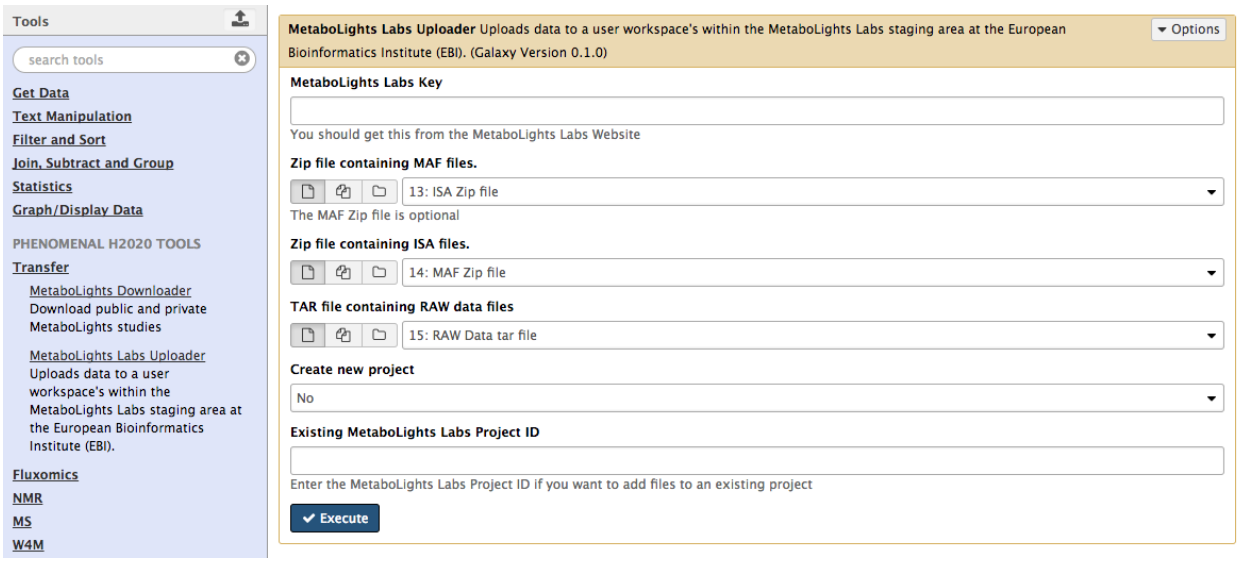


**Figure 5:** Screenshot of the MetaboLights Labs Galaxy wrapper on the PhenoMeNal public instance Galaxy VRE. The user can select data from files available in their Galaxy history for upload. When uploading, the user can either create a new Project within the MetaboLights Labs Workspace or provide the Project ID for an existing Labs Project to add more data. The user is identified through a personal MetaboLights user API Key, which is obtained in the users personal profile in MetaboLights (the user needs to be registered with MetaboLights). To the left of the figure, in the purple area, the user can find the MetaboLights Lab uploader under the section “Transfer”.

The Galaxy MetaboLights Labs Uploader module receives as input a set of files (Zip file containing MetaboLights Metabolite Annotation Files, Zip file containing ISA-Tab documents and TAR file containing RAW data files), the MetaboLights Labs API Key, whether to create a new MetaboLights Labs Project and, if desired, an existing MetaboLights Labs Project ID to add data to.

The Galaxy wrapper described here calls a docker container built to encapsulate the Python utility, available through:

```
docker pull docker-registry.phenomenal-h2020.eu/phnmnl/mtbl-labs-uploader
```

for the actual upload process. The Python wrapper described resides inside the docker container and can be invoked (on a machine running docker) with:

```
docker run -w=$PWD -v /home/your_user:/home/your_user \ docker-registry.phenomenal-
h2020.eu/phnmnl/mtbl-labs-uploader \
-t <your_metabolights_key> \
-i path/to/metabolomics_data.tar path/to/isa.zip -n
```

### 3.6. Study subsetting and data volume transfer considerations

When downloading whole datasets from MetaboLights, the entire set of raw files can be rather large, so transferring only parts of a study that are to be used from MetaboLights is preferable. As mentioned in previous Sections, PhenoMeNal VRE users can selectively access portions of datasets and associated raw files based on experimental factors defined within the metadata files (ISA-Tab), for example gender or healthy vs disease.

#### 3.6.1. ISA slicer

The ISA API[16] is a Python 3 library that can create, manipulate, and convert ISA formatted content. The mtbls.py module[17] within the ISA API provides the functionality to access MetaboLights ISA-Tab data, wrapped up as the mtblisa container[18]. Specifically it allows one to:

- Retrieve metadata from MetaboLights studies in the ISA-Tab and JSON formats
- Query MetaboLights studies for factors used.
- Query MetaboLights studies for factor values used for a given factor.
- Query MetaboLights studies to retrieve data file names filtered on factor and factor value.

What this allows us to do is to get a subset of data file names for a particular MetaboLights study of interest, without download the raw files first, since all of the study's data files are listed in the ISA-Tab metadata. This provides us with a neat way of selecting our data files of interest without downloading the whole study. File URLs can then be constructed to use within the MetaboLights downloader, mtbls-downloader, described in previous sections.

### 3.7. File conversion and metadata extraction

Raw data in vendor formats encodes not only the spectral data, but also some metadata and instrument settings, several of which can also be encoded in the open mzML format. For the purpose of this document we are focusing on ProteoWizard for the conversion of a growing number of Mass Spec data formats to the open mzML format.

[16] https://github.com/ISA-tools/isa-api/
[17] https://github.com/phnmnl/isa-api/blob/5cc3c9efa8dfff498e3e421ae930a3e471f28229/isatools/io/mtbls.py
[18] https://github.com/phnmnl/container-mtblisa

### 3.7.1. ProteoWizard converts to mzML

ProteoWizard is a software package that runs only on Windows. The reason for this limitation is a set of vendor-specific libraries (DLLs) that are only available on Windows. To use it within the PhenoMeNal platform, we run this tool in a Docker container using the open source Wine[19] (Windows Emulator) software stack. Wine implements Windows API calls on UNIX/Linux systems, so one can run Windows software on UNIX/Linux. The ProteoWizard containerisation was described in Deliverable 9.2.1. We are currently using ProteoWizard to support the mzXML and Bruker "*.d" file formats. Support for other vendor data formats will be added during the project.

### 3.7.2. Mzml2isa Module extracts metadata and instrument settings

Mzml2isa is a program that automatically generates an ISA-Tab document structure from raw XML metabolomics data files (mzML open access data format). The mzml2isa tool provides the initial structure for the ISA-Tab documents required for a MetaboLights study.

The ISA-Tab format is a set of tab delimited spreadsheet-like files that describe the *Investigation*, one or more *Studies*, and one or more *Assays* per study. The `i_Investigation.txt` file captures the title and brief description of the underlying aim of a biomedical investigation, including references to the used ontologies, a list of protocols applied, bibliographic information and contact data.

The `s_Study.txt` file(s) describe the origin of the sample material, characteristics, protocols and experimental design factors relevant to the individual samples. The `a_Assay.txt` file(s) for metabolomics assays contains information regarding how the individual samples were extracted, possibly labelled and the analytical protocols and their parameters used for the actual measurements. The ISAcreator metabolomics plugin developed at the EMBL-EBI captures the reporting of metabolites measured as Metabolite Assignment File (MAF).

The metadata generated by mzml2isa can be further enhanced in MetaboLights Labs for a final study publication. We created the docker container for mzml2isa – the software tool and corresponding Galaxy module were already available for the community (at github.com/althonos/mzml2isa and github.com/ISA-tools/mzml2isa-galaxy respectively). Incorporating the tool in a container allows it to run within our container orchestrator cluster environment/deployment. The new container definition is available at github.com/phnmnl/container-mzml2isa/tree/develop. The mzml2isa tool and Galaxy module available received extensive testing within our infrastructure,

[19] https://www.winehq.org/

resulting in the contribution of several bug fixes by PhenoMeNal developers to the original projects through two issues and four pull requests (#1, #2, #4 and #5) respectively. After all these bug fixes and improvements, the mzml2isa Galaxy module and container were added to the PhenoMeNal public instance and Galaxy deployment scripts.

The mzml2isa module takes as input a Zip or TAR file bundle of mzml files and a plethora of metadata that the user can input through the Galaxy UI (as Figure 6 shows), to produce ISA and MAF files that contain both metadata entered by the user as well as obtained/inferred from the provided mzml files.

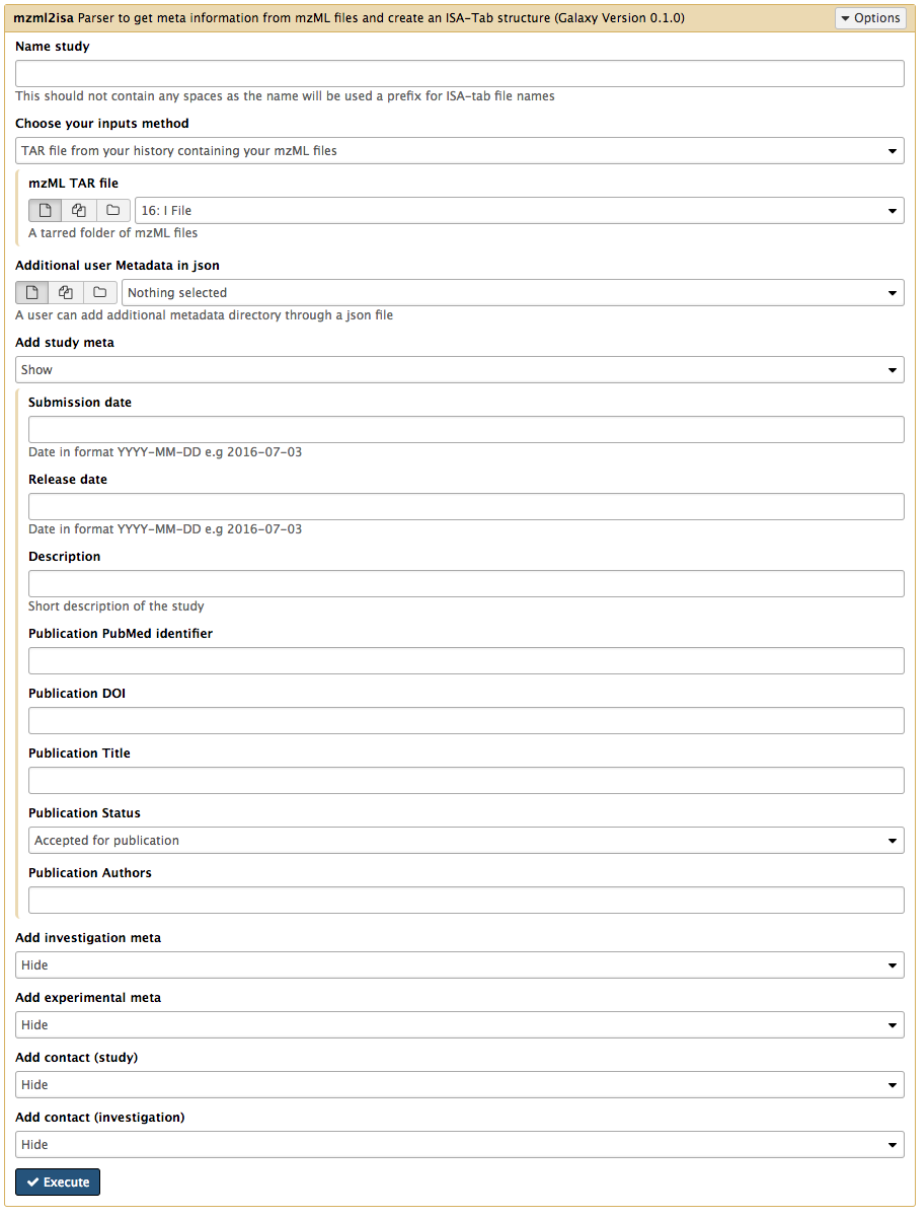


**Figure 6:** Metadata input available for mzml2isa through the Galaxy UI that the module presents to the user. Additional structure metadata can be supplied through a JSON file. The user can manually input study metadata (shown expanded), investigation metadata, experimental metadata, contact data and investigation metadata. Additional metadata is also harvested from the mzML files provided. The resulting ISA and MAF files point to these mzML files. These files form the basis for a possible data submission to MetaboLights. MetaboLights support ISA-Tab documents for metadata annotations.

# 4. WORK PLAN

The work was planned and followed using Pivotal Tracker[20]. The following were the main stories and have all been completed:

| Story ID | Story Name |
|---|---|
| 130515937 | Create container for mzml2isa |
| 131723417 | mzml2isa galaxy wrapper working on PhenoMeNal |
| 131716035 | Dockerise Aspera client for VRE |
| 131715653 | MetaboLights to PhenoMeNal VRE raw data link - D9.2.2 |
| 131715937 | Python wrapper to access MetaboLights Labs web service from Galaxy |
| 131715205 | Galaxy Wrapper to upload data to MetaboLights Labs - D9.2.2 |
| 133025351 | MetaboLights LABS -Study folder synchronisation jobs |
| 133025775 | Extend MetaboLights LABS web services |
| 131715489 | PhenoMeNal to MetaboLights data dissemination - D9.2.2 |
| 126671539 | Script for pulling from MTBLS for use in ISA API |
| 133360467 | ISA slicer |
| 133360525 | Containerize ISA slicer |

## 4.1. Utilisation of resources:

Person Month (PM) contribution towards this deliverable:

| **Partner** | **EMBL-EBI** | **ICL** | **IPB** | **UB** | **UOXF** | **UU** | **CEA** | **INRA** | **CRS4** |
|---|---|---|---|---|---|---|---|---|---|
| PMs | 2 | 1 | 2 | 0.5 | 3 | 2 | 2 | 3 | 0.5 |

[20] http://phenomenal-h2020.eu/home/about/project-management-tool/

## 5. DELIVERY AND SCHEDULE

The deliverable is submitted on time.

## 6. CONCLUSION

In this deliverable, we described the data transfer and management architecture in PhenoMeNal. Data can be uploaded either directly to a PhenoMeNal Galaxy VRE for analysis. In addition, we have created additional tools within Galaxy, to perform a direct data transfer between MetaboLights and MetaboLights LABS, thus avoiding slow and duplicated up- and downloads through the user's web browser. The data transfer is accelerated using the highly efficient Aspera network protocol. Other data transfer mechanisms are available as well, to avoid any vendor lock-in.

The metadata is structured in the ISA-Tab format, which describes the Investigation, Studies and Assays. The format is supported by a large number of academic and commercial sample management systems, programming frameworks and database not only in Metabolomics. Users can subset data by experimental design factors and phenotypic characteristics, allowing dedicated analysis but also the reduction of data transfer requirements.

# 7. ANNEX

## 7.1. Aspera compare

Example of real world timings for MTBLS20 (4.5 GB):

| Server/Protocol | Download speed / Time elapsed |
|---|---|
| **FTP server (individual files):**<br>$ time wget ftp://ftp.ebi.ac.uk/pub/databases/metabolights/studies/public/MTBLS20/* | 25 Mbps: 27m 20s<br>850Mbps: 1m 49s |
| **Aspera server (individual files):**<br>$ time ascp -QT -l 1g fasp-ml@fasp.ebi.ac.uk:/studies/public/MTBLS20 . | 25 Mbps: 25m 36s<br>850 Mbps: 0m 50s |