Deliverable 4.1

Project ID	654241
Project Title	A comprehensive and standardised e-infrastructure for analysing medical metabolic phenotype data.
Project Acronym	PhenoMeNal
Start Date of the Project	1st September 2015
Duration of the Project	36 Months
Work Package Number	4
Work Package Title	Interfacing with Biomedical European Infrastructures
Deliverable Title	D4.1 Report on requirements for relevant research centres producing and/or consuming metabolomics data with respect to computational aspects, data storage, and infrastructural needs
Delivery Date	M12
Work Package leader	CIRMMP
Contributing Partners	CIRMMP, ICL, UB, UOXF
Authors	Antonio Rosato
Abstract: This deliverable summarizes the requirements on various IT aspects expressed by research centres producing and/or consuming metabolomics (and other -omics) data. It will be updated as more information is received.	

Table of Contents

1. EXECUTIVE SUMMARY

The activities of WP4 aim to foster the interactions of PhenoMeNal with other research centres (public or private), large infrastructures (such as ESFRIs) and national, regional or European initiatives that produce or consume metabolomics data as part of their routine work. The rationale is that such centres and infrastructures constitute a relevant part of the potential user basis of PhenoMeNal. In the future, they may also become partners of spin-off initiatives from the present project. Therefore, it is important that the services offered by PhenoMeNal are aligned with their requirements, so that the usefulness of such services can be readily perceived.

Within the context outlined in the previous paragraph, we have contacted a number of major centres/infrastructures that produce and/or consume metabolomics data, including but not limited to those listed in the mapping of deliverable **D2.1** (Report on mapping of e-infrastructures, users, investments for supporting policy developments in the field of metabolomics, biomerkers and biobanks, submitted M6). Our aim is to collect information on their computational and storage requirements, as well as on some aspects of software usage. This information will help the PhenoMeNal consortium to produce an infrastructure that provides tools effectively useful to the community.

The following general observations have been made

- European centres are currently not making systematic use of either cloud computing or cloud storage.
- Consequently, they exploit local computational infrastructures. The majority of centres fit within a computing facility of 1,000-2,000 cores with 8 GB memory per core.
- There are two different sets of storage requirements: nearly half of the sites do not clearly separate short- and long-term storage, and require space up a few tens TB. The other half of the respondents have a defined policy for data storage, with several tens TB for short-term storage and up to a few PB (nearline disks or tapes) for long-term storage.
- Distributed filesystems are in use only at the largest sites.
- Data throughput is quite diversified. The majority of the centres have monthly throughputs of the order of 1 TB or less. The forecasts on the future growth of such throughput are around 100 TB per year, up to 1-2 PB per year.

- Software requirements are the requirements for which there is most consensus among respondents. Nearly all use MySQL and/or PostgreSQL as their database management systems. The majority of the centres using workflow managers have adopted Galaxy[1]. Statistical analysis is universally based on R[2], with some use of Matlab and Python (e.g. pandas). The most commonly used graphical interface is RStudio[3], in addition to vendor-specific software.
- Finally, most respondents did not address issues such as Big Data analytics, PaaS (Platform as a service) and provisioning tools.

PhenoMeNal is well in line to provide usable solutions to several of the points above. As reported in deliverable **D5.2** (A beta-version of PhenoMeNal integration VMI capable of proof-of-concept integration with other VMIs. Initial services online supporting PhenoMeNal data standards, submitted M12), PhenoMeNal can already handle Galaxy workflows and has successfully demonstrated an R-based metabolomics workflow. The integration of PostgreSQL is planned.

2. CONTRIBUTION TOWARDS PROJECT OBJECTIVES

The activities described in the present report contribute to the achievement of the following objectives:

Objective 4.1 Boost the offering of services by the PhenoMeNal e-infrastructure to the current large-scale EU biomedical infrastructures and national and regional research centers, and their users.
Objective 4.2 Align PhenoMeNal activities to the requirements of such infrastructures and centers.

Largely, they contribute towards the first two general objectives of the project :
Project Objective 1: To integrate exsisting open source tools and methods for the management, dissemination and computational analysis of very large datasets of human metabolic phenotyping and genomic data into a secure and sustainable e-infrastructure
Project Objective 2: To operate and consolidate the PhenoMeNal e-infrastructure based on existing internal and external HPC and grid resources, including the EGI, and to extend it to world-wide grid infrastructures.

[1] https://usegalaxy.org
[2] https://www.r-project.org
[3] https://www.rstudio.com/home/

3. DETAILED REPORT ON THE DELIVERABLE

3.1. Preparation of a questionnaire to collect requirements

We decided that the easiest way to collect the needed feedback was through the preparation and distribution of a specific questionnaire. This questionnaire was intended to allow the staff in charge of IT and bioinformatics at each centre/infrastructure to give us information on selected aspects relevant for the development of the services of PhenoMeNal. This approach was chosen because in the context of WP4 the goal is to reach out to centres as a whole rather than to the individual researchers working there.

The questionnaire was agreed upon through various rounds of discussion involving all partners using the different communication tools of the project: hangouts, Slack, and PivotalTracker (online system of project management).

In its final version, the questionnaire addressed the following key topics:

a. Type of e-infrastructure
b. Computational requirements
c. Storage requirements
d. Data throughput
e. Software requirements

In addition to computational, storage and software requirements, we inquired about the e-infrastructure already in use (type, size, etc.) as well as the current and foreseen data throughput. It is particularly relevant to gather information about data throughput because of the focus of PhenoMeNal on high-throughput processing and analysis of omics data. The full questionnaire can be found under **annex section 7.1** . All partners provided contact to send the questionnaire to and/or circulated the questionnaire amongst themselves. To facilitate the respondents, we provided both a doc file and an online feedback via SurveyMonkey[4].

[4] https://www.surveymonkey.com/r/YNTVTCJ

3.2. Analysis of feedback

We received a compiled questionnaire, either via the online survey or via email as an attachment, from 15 respondents who were either individual centres or hubs of centres for a total of 18 sites, across Europe. One additional questionnaire was received from TMIC (The Metabolomics Innovation Centre), which is based in Canada. Not all sites identified themselves in the online survey. Because the survey was distributed after mid-June, there has been a slow response from some contacts, due to holidays. We decided to leave it open past the due date of D4.1 to continue to receive input and then update this report after M12.

According to our design, nearly all the responses were from relatively large centres, public or private, with a significant vocation toward service provision. About 75% of the respondents specialized in metabolomics, whereas the remaining 25% featured multiple omics platforms. The majority of the metabolomics sites used mass spectrometry (MS) or MS together with NMR spectroscopy (NMR). Only two sites used NMR exclusively. The fields of application encompassed all aspects of metabolomics in biomedicine, including target and untargeted metabolomics, personalized medicine, lipidomics and fluxomics. Service provision focused mostly on experimental work but included also data processing and analysis (particularly for the multi-omics platforms).

A summary of the responses to the questionnaire is as follows:

Type of e-infrastructure:

The typical hardware setup involves mainly the use of CPU clusters, with some additional usage of HPC resources. Distributed (grid) computing is in use only at one centre. The queuing systems in use are quite diverse, with TORQUE (Tetrascale Open-source Resource QUEue Manager) and Grid Engine being the two most common (but three other solutions have been mentioned). Public clouds are not used by the European centres, with the exception of EBI; two European respondents have done some testing or had experience for very specific project-driven purposes. On the contrary, TMIC e-infrastructure is mainly based on public cloud computing (Google Cloud and Digital Ocean). Private clouds based on OpenStack are in use at two European sites and TMIC. The use of big data analytics is also very restricted (Apache Hadoop at only one site), as is the use of PaaS (only for LifeRay).

Computational requirements:

A large majority of sites uses between a few hundreds and a few thousands of CPU cores. Two sites additionally have small (6 and 18 cards) GPU clusters. The memory available per CPU core is very variable, between 2 and 128 GB. TMIC has a local infrastructure of about 180 cores, with an average of 2.5 GB of memory per core. Overall, it appears that most sites would be satisfied with 8 GB per core. No sites have a requirement for a minimum number of cores on a single processor. Finally, provisioning tools are not commonly used: two sites use Ansible[5], one site uses Vagrant[6] to instantiate VMs and TMIC uses SaltStack[7] and Capistrano[8].

Storage requirements:

The storage requirements are very diversified. About half of the sites have separate long-term and short-term storage space, with some specific policies to move data from one to the other. These sites use larger storage than the sites that do not have such policies in place, ranging from several tens up to a few hundreds of TB for the short-term storage and a few PB for the long-term storage (if on disk, alternatively tapes are used). The sites that do not separate long-term and short-term storage space use up to only 10-20 TB of storage. There is no strong correlation between the storage and computational requirements, in that even sites that use little storage space can have a computing capacity of the order of 1,000 cores. All storage is on-site: cloud storage is not used by the respondents, except TMIC that is using a commercial DropBox automated backup (10 TB). The use of distributed filesystems is modest: only the two largest sites (UK Phenome Centre and CRS4[9]) use iRods (integrated Rule-Oriented Data System), whereas TMIC uses Samba. One site has adopted GlusterFS. It is however possible that the use of distributed filesystems is “hidden” because the storage infrastructure is handled by others, such as the IT department of the local University.

Data throughput (here defined as the data generated by each center per month):

8 of the respondents have a throughput lower than 1 TB per month, 3 have a throughput between 1 and 5 TB per month,1 has a throughput of 10-20 TB per month, and one (the UK National Phenome Center) has a throughput of 120 TB per month. The forecast increase of this throughput over the next few years is very variable (from steady up to a

[5] https://www.ansible.com
[6] https://www.vagrantup.com
[7] https://saltstack.com
[8] http://capistranorb.com
[9] http://www.crs4.it

10-fold increase). The highest predictions are of a throughput of 1-2 PB per year (i.e. 100-150 TB per month, similar to the current production of the UK National Phenome Center). Regarding the protection of sensitive data, the approaches implemented are obviously largely dependent on applicable regulations. The two most common lines are data pseudoanonymization or the restriction to local access only, together with access control.

Software requirements:

Are those with the greatest similarity among the respondents. Essentially all respondents that use a database management system adopted MySQL and/or PostgreSQL, with two exceptions (Microsoft Access and Microsoft Excel). The most common workflow management system is Galaxy, which is exploited by all but two of the respondents using WMS. Alternatives are gUSE (grid and cloud User Support Environment), or in-house pipelines. The use of APIs instead is not very widespread; some include BioRuby, BioPerl, and BioJava. All respondents except one use R; Matlab and Python tools/libraries for data analysis (such as Python pandas) are also quite common. Accordingly, the most commonly used graphical interface is RStudio; also MetaboAnalyst[10] is in use at more than one site. Several sites additionally use a variety of vendor-specific software.

PhenoMeNal is already well positioned to address several of the software requirements outlined above. A solution based on connecting Galaxy to Kubernetes has been achieved by means of the Galaxy Kubernetes Runner (deliverable **D5.2**). This allows Galaxy to build workflows that use the tools that are pre-packaged as containers, in **WP9** (Tools, workflows, audit and data management) or by external users, on top of a container orchestrator cluster (Kubernetes). The container orchestrator cluster can in turn be deployed to an elastic cloud infrastructure or to local machines. This infrastructure can go from a just a few nodes (even one) to thousands of nodes if required, with variable number of cores, making no difference in terms of how our Galaxy instance and tool deployment is done on top of the container orchestration. This allows to cater for the wide range of hardware specs detailed in the requirements above. In terms of multi-threaded operation of tools, the container orchestrator based installations poses no a-priori limitation, and will only depend on the amount of cores and core per memory supplied in either the VM appliances or real hardware where the container orchestrator (including our Galaxy instance and tools) is deployed. The infrastructure is normally deployed with a scalable filesystem (currently GlusterFS), which allows prospective users to increase capacity and performance as they see fit (by adding more nodes). Changes in the filesystem layer capacity/performance again make

[10] http://www.metaboanalyst.ca

no difference to the Galaxy+tools installation on top of the container orchestrator, as this layer is abstracted by Kubernetes to the running processes. In parallel, our proof of concept demonstrators included an R-based metabolomics workflow. Some Galaxy-based metabolomics workflows have also been implemented as part of the demonstrators. These metabolomics workflows are available at our PhenoMeNal public Galaxy-Kubernetes instance at http://public.phenomenal-h2020.eu/, using the technologies described. The current work plan also includes introducing the capability to use PostgreSQL as backend for Galaxy, improving HTML cache, reducing Galaxy's and tool's docker container image sizes and migrate more tools to be able to satisfy all of our use cases. Future work plans include the provisioning of a public instance of Jupyter/iPython to work with our container-based tools on top of the container orchestrator layer (either through the Galaxy API or directly, to be decided), as is the case currently with Galaxy through the Kubernetes Runner. We aim to run tests for compute/memory intensive tools, as the one done for IPO (**Deliverable D5.2**), to understand the cores to memory relation for them. Tests on federated data access will start soon as well, to try systems like iRods and Ubernetes (federated Kubernetes), which might require adaptations at the Galaxy layer as well.

3.3. Feedback from industry

On the occasion of the PhenoMeNal industry workshop organized on Tuesday June 28th at the Dublin Conference Centre by **WP2** (Sustainability of PhenoMeNal), a simplified version of the questionnaire was distributed to the industry invitees. The specific question was to provide us feedback on the technical requirements for PhenoMeNal to be optimally interfacing with their technologies. We received feedback from three of them (Agilent, Bruker, Biocrates). Their input is briefly summarized below.
As one might anticipate, vendors and metabolomics service providers crucially rely on their own solutions also for data processing and analysis. For this, they have developed integrated workflows whose source code is not publicly available. Such workflows are distributed to customers as closed packages and require local installation. There are different approaches to providing APIs to the workflows, including licensing APIs or the use of standard REST/SOAP/CORBA protocols.
The interaction of the PhenoMeNal infrastructure with industry workflows should thus focus on facilitating data flows among these different resources. In this scenario, the metadata management capabilities developed by PhenoMeNal appear of particular relevance.

4. WORK PLAN

4.1. Structure and management of WP4 tasks

The aim of WP4 is to maximize the interaction of PhenoMeNal with European infrastructures and national or regional research centres with an interest in biomedical data generation and analysis. In this way, the present consortium can inform other infrastructures and research centres about the development of the PhenoMeNal e-infrastructure, remaining aligned with the progress in the field and the needs of PhenoMeNal potential users.

The present deliverable is part of task T4.1

Task 4.1: **Collecting and reporting on the requirements** of research and/or industry centres those produce or make use of metabolomics (possibly together with other –omics techniques) data (CIRMMP, ICL, UB, UOXF, UL, EMBL-EBI, UU). In particular, we will evaluate computational, storage and software requirements based on current and forecast data throughput.

The following objectives are targeted:

- **Objective 4.1** Boost the offering of services by the PhenoMeNal e-infrastructure to the current large-scale EU biomedical infrastructures and national and regional research centres, and their users.
- **Objective 4.2** Align PhenoMeNal activities to the requirements of such infrastructures and centres.

WP4 is led by CIRMMP who is responsible for planning and coordinating the work and the related deliverables. The planning was carried out in collaboration with other partners; ICL, UB, UOXF, UL, EMBL-EBI, and UU. UL asked for specific requirements by instrument vendors on the occasion of the Industry meeting held in Dublin. The input of all partners was collected to finalize the deliverable and obtain contacts for the distribution of the questionnaire. The progress was tracked using Pivotal Tracker (see Figure1)

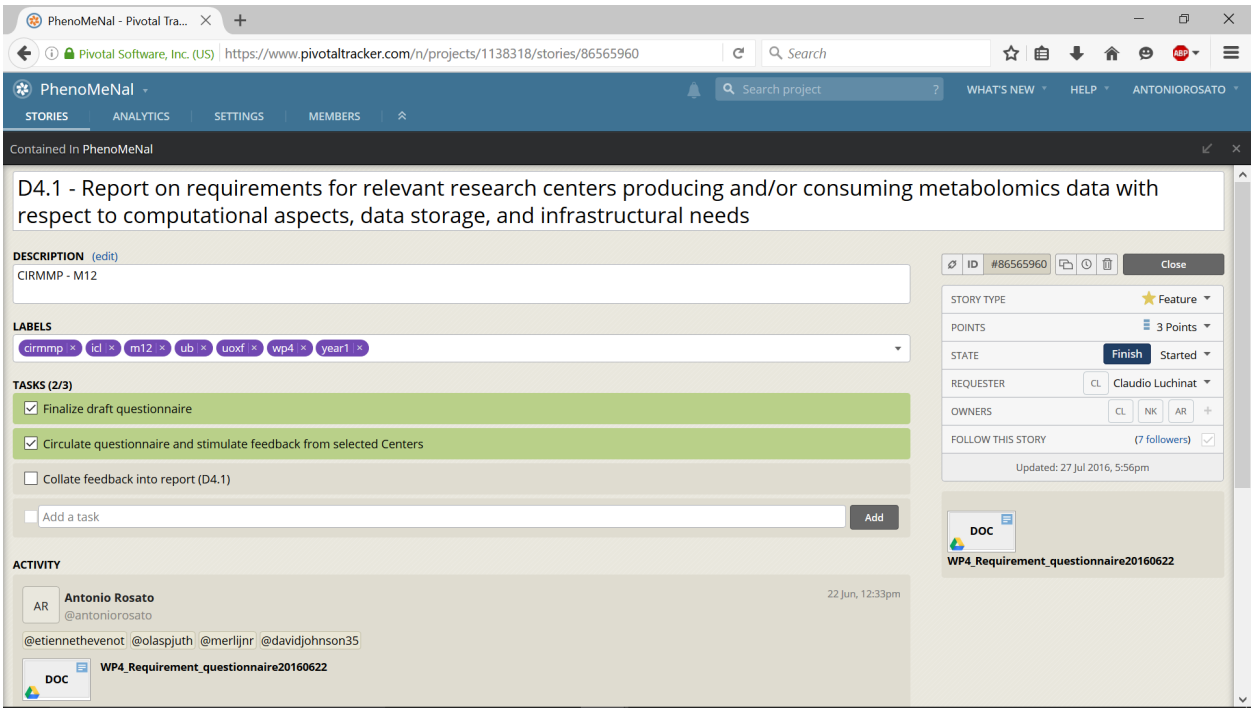

Figure 1. Description and breaking of tasks in Pivotal tracker

Utilization of resources:

The total PMs (person months) utilised until M12 (inclusive):

Partner	CIRMMP	ICL	UB	UOXF	UU
PM	7,5	0.1	2	2	1

5. DELIVERY AND SCHEDULE

The deliverable was submitted on time.

6. CONCLUSION

The questionnaire to collect requirements by infrastructures and large research centres generating and/or consuming metabolomics (and other omics) data was circulated to a number of institutions. Enough feedback was received by mid August to obtain an

informative overview on numerous aspects of the IT infrastructure at these centres. Such overview is useful to align PhenoMeNal services to the actual needs of this part of its user community (i.e. not the individual researchers but experimental infrastructures as a whole).

- European centres are not making systematic use of either cloud computing or cloud storage
- Consequently, they have local computational infrastructures. The majority of centres fit within a computing facility of 1,000-2,000 cores with 8 GB memory per core
- There are two different sets of storage requirements: nearly half of the sites do not clearly separate short- and long-term storage, and require space up a few tens TB. The other half of the respondents have a defined policy for data storage, with several tens TB for short-term storage and up to a few PB for long-term storage
- Distributed filesystems are in use at the largest sites
- Data throughput is diversified. The majority of the centres have monthly throughputs of the order of 1 TB or less. The forecasts on the future growth of such throughput are around 100 TB per year, up to 1-2 PB per year.
- Software requirements are the requirements for which there is most consensus among respondents. Nearly all use MySQL and/or PostgreSQL as their database management systems. The majority of the centres using workflow managers have adopted Galaxy. Statistical analysis is universally based on R, with some use of Matlab and Python pandas. The most commonly used graphical interface is RStudio, in addition to vendor-specific software. Some of these tools are already being implemented in PhenoMeNal (see deliverable D5.2).
- Finally, most respondents did not address issues such as Big Data analytics, PaaS and provisioning tools

It is interesting to use the above data to estimate the cost of implementing the “average” computational infrastructure on a public cloud such as AWS. Assuming 1000 two-core instances per month with 8 GB memory per instance, running 100% of time, results in a cost of about 90 k$ per month for the compute, while data transfer and temporary storage do not contribute to the cost. This cost is about halved if the memory is reduced to 4 GB.

7. ANNEX

7.1. PhenoMeNal questionnaire

QUESTIONS TO COLLECT REQUIREMENTS FROM
NATIONAL AND REGIONAL INFRASTRUCTURES, ESFRI's, OTHER (LARGE) CENTERS

The expressed requirements can be actual or anticipated.

Brief (five lines) description of the area of work/mission of the infrastructure (Specify relevance of metabolomics)

Type of e-infrastructure
Please explain the e-infrastructure types required/used:

- High-performance batch system; please specify queueing system
- Public (e.g. Amazon EC2, Google cloud, Microsoft Azure) and/or Private cloud resources (Infrastructure-as-a-Service, e.g. OpenStack)
- Big Data analytics (e.g. Apache Hadoop, Spark, Flink)
- Platform-as-a-Service (specify which platforms)

Computational requirements
For each in-house computer system, please indicate

- Number of CPU cores used/needed for data analysis
- Is there a requirement for processors with a minimal number of cores. If yes, how many
- Memory per CPU core/processor
- Provisioning tools used (Terraform, Ansible, Puppet, ...), if any?

Storage requirements

- Are there separate short-term/long-term storage systems? If yes, is there a policy to move data from one to the other?
- How much total storage space is available for each system?
- Is external data storage required/used (e.g. public cloud)?
- Any experience with distributed filesystems (like iRods)?

Data throughput

- What are the typical amount of data generated (e.g. monthly average) and file/dataset size?
- If external data storage is in use, how much data are transferred to/from it?
- If the data contain sensitive information, what security measures are in place or required?
- Do you have an estimate of the expected growth of the data volume in the next five years?

Software requirements

- Database management system (please specify)
- Workflow management system (please specify)
- Statistical environments (please specify)
- Programming interfaces, APIs (please specify)
- Graphical interfaces/portals (please specify)