

Deliverable 2.5

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	2
Work Package Title	Sustainability of PhenoMeNal
Deliverable Title	D2.5 Report on concrete sustainability measures taken by all partners
Delivery Date	M32
Work Package leader	UL
Contributing Partners	All
Authors	Namrata Kale, Merlijn van Rijswijk, Antonio Rosato, Marta Cascante, Pedro De Atauri, Vitaly Selivanov, Daniel Schober, Steffen Neumann, Ken Haug, David Johnson, Philippe Rocca-Serra, Luca Pireddu, Ralf Weber, James Bradbury, Tim Ebbels
Abstract	This deliverable describes the measures taken by the consortium for the sustainability of PhenoMeNal beyond its 3-year funding period.



Contents

1	Executive Summary	3
2	Contribution towards the project objectives	3
3	PhenoMeNal sustainability plan	3
3.1	Economic sustainability	4
3.1.1	PhenoMeNal Foundation	5
3.1.2	User community engagement	7
3.1.3	Stakeholder engagement.....	13
3.2	Technical Sustainability	17
3.2.1	Security	17
3.2.2	PhenoMeNal VRE portal.....	18
3.2.3	Sustainability measures for (individual) tools and containers	19
3.2.4	Data Formats and Terminologies	26
3.2.5	Workflow engines	27
3.2.6	Deployment of PhenoMeNal VREs	27
4	Delivery and Schedule	29
5	Conclusion.....	29
6	Annex 1	31



1 Executive Summary

Metabolomics, the data-intensive study of all metabolic processes, is a rapidly advancing field with a diversity of applications across biomedical domains, including personalised medicine. Metabolomics opens up exciting opportunities for research, but is grappling with the technical challenges posed by very large data volumes. The PhenoMeNal project offers a solution by providing a comprehensive, standardised e-infrastructure that supports data processing and analysis in cloud environments. The tools developed by the project are open-source and the platform can be applied to other scientific domains, including other -omics studies. The project thus operates in a complex bioinformatics ecosystem of research infrastructures in biomedicine and systems biology, as well as other e-infrastructures, which together provide cloud computing, data services, biological samples and medical technologies. Broad acceptance by the community and developing services that are well supported are thus the key sustainability measures adopted by the consortium.

This document describes the concrete sustainability measures taken by all partners to ensure the widespread use, continuity and further development of the tools and services offered by PhenoMeNal after the end of the funding period.

2 Contribution towards the project objectives

The deliverable has contributed towards the following project objectives for WP2:

- Mapping of e-infrastructures, users, investments for supporting policy developments in the field of metabolomics, biomarkers and biobanks.
- Establishing and maintaining relations with publishers for supporting data deposition services.
- Coordination with national and international related activities and support to the development of global and sustainable approaches in the field.
- Promotion of long term sustainability, including the involvement of funders.
- Preparation of business plan beyond the end of the project.

3 PhenoMeNal sustainability plan

Since the inception, the consortium has worked towards developing and implementing plans to advertise and sustain the PhenoMeNal infrastructure beyond its 3-year funding



period. The overall sustainability plan¹ (see figure 1) covers **technical sustainability** as well as business model options for **economic sustainability** of the PhenoMeNal infrastructure.

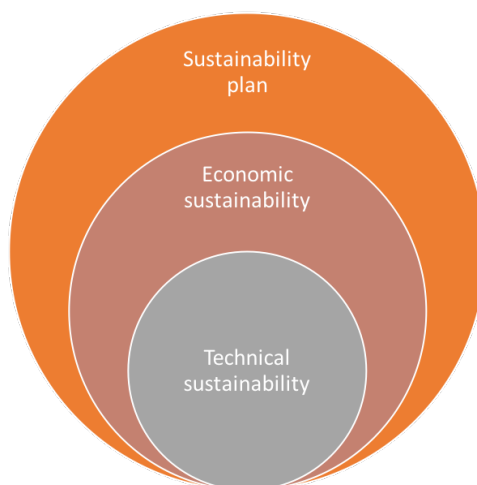


Figure 1. PhenoMeNal sustainability plan

3.1 Economic sustainability

As suggested in its initial sustainability plan (D2.2), the consortium has pursued towards creating an “**Active PhenoMeNal community Foundation**”. PhenoMeNal is a new e-infrastructure and as shown in figure 2, the most relevant element for the sustainability of PhenoMeNal is a) its broad acceptance and involvement of users, b) establishment of good links with other e-infrastructure initiatives (like ELIXIR²) and research infrastructures in biomedicine (like BBMRI³ and [ISBE](http://project.isbe.eu/)⁴). Significant efforts have therefore been made to test the different components of the PhenoMeNal infrastructure and gather inputs from users regarding their requirements – both from “expert” users, like bioinformaticians, as well as “knowledgeable” end users such as clinicians and epidemiologists.

¹ <http://phenomenal-h2020.eu/home/wp-content/uploads/2016/09/Deliverable2.2.pdf>

² <https://www.elixir-europe.org>

³ <http://www.bbmri-eric.eu>

⁴ <http://project.isbe.eu/>

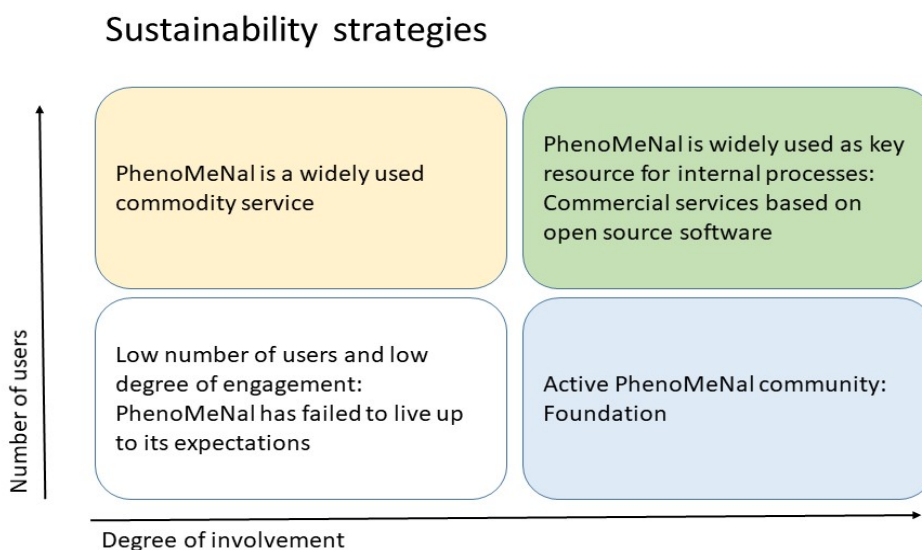


Figure 2. Sustainability strategies for PhenoMeNal

3.1.1 PhenoMeNal Foundation

The PhenoMeNal consortium has unanimously voted for the creation of a *foundation*, for the sustainability of PhenoMeNal after the end of the project. The proposal for foundation has largely been inspired by the already established tranSMART foundation for precision medicine⁵.

This initiative is led by University of Leiden and work is currently in progress on the rules of this foundation with a notary, which will be open for feedback by the consortium once the first draft is finalised. The setting up the foundation has been delayed (scheduled for M30 in the report D2.4 Business plan beyond the end of the project, submitted in M24), following the decision to involve the legal departments of consortium partners in the preparation of the general outline of the foundation. This process has been proven to be useful and constructive, yet more time-consuming than anticipated.

The overall aim of the foundation is to create a sustainable administrative/management entity that will steer and control all aspects of the PhenoMeNal service portfolio, and align these with other community initiatives (for instance the GO FAIR Implementation network and the ELIXIR Metabolomics Community, as described below). The consortium has operationalised this in the following subaims:

⁵ <http://transmartfoundation.org/about-the-pta/>



- Steer the maintenance and control the further development of PhenoMeNal.
- Organize a community of users and developers and foster the interactions.
- Generate funding, by applying for grants (nationally and internationally) and by aiming to negotiate deals with relevant users of the PhenoMeNal infrastructure (for instance Phenome Centre Network, Academic Medical Centers, Instrument Vendors, companies applying Metabolomics).
- Organize trainings and help for users.
- Outreach to new users, for instance via a website.
- Support the development of metabolomics computational tools in general, and fostering the development of standards and guidelines and their implementation.

The aims of the foundation on purpose have been formulated broadly, in order to prevent that modifications would be needed, to support any new activities of the foundation in the near future.

Way of working of the Foundation

If a partner wishes to join the PhenoMeNal Foundation, a **partnership agreement** will be drafted between the Foundation and the partner. In this agreement, the rights between the partner and the foundation are arranged, for instance:

- The partner signs to **underwrite the aims of the foundation, and to invest from their own means (in-kind) into sustaining the PhenoMeNal infrastructure** (to the best of the partner's ability, no "hard" legally binding commitment)
- The partner agrees to pay an annual **partnership fee**. This fee can be waived by the Board, if the partner invests considerable means towards sustaining the infrastructure.
- An obligation of a **minimum duration of the partnership** (proposal: minimal two years, thereafter this can be ended every year, with a minimum notice of 1 months before the end of the yearly period). An automatic renewal will be in place.
- A right to **nominate one person for the "general assembly"** of the foundation, which meets at least once a year

Governance

The foundation will have a Board, which consists of at least 5 - 9 members (to be manageable). It is expected that key contributors of PhenoMeNal will take a Board position. In addition, if an organisation becomes a large supporter of PhenoMeNal (instrument vendors, academic medical hospital, a pharma or nutritional company



applying metabolomics) in reward also a Board position can be offered (to represent the interest of similar ‘stakeholders’ in the Board). All Board members will be registered at the Chamber of Commerce. Board members will take a seat in the Board on a personal basis (meaning, they are not representing their own organisations) for a period of three years.

3.1.2 User community engagement

The consortium has concentrated its efforts in building a strong relationship with the user community for a high and lasting impact. It has ensured that PhenoMeNal services are user-centric and adapted to the needs of both the early-stage and the advanced research community.

3.1.2.1 User designed VRC portal

The PhenoMeNal Virtual Research Community (VRC) portal, also known as the Cloud Research Environment (CRE) portal⁶, is the central access point to interact with and accessing the compute resources for data analysis. Through iterative usability testing^{7,8}, the portal has been implemented using modern web technologies and design guidelines to provide the best user experience, maximising the users ability to accomplish the required tasks. For different stages of the UX testing, see figure 3 below.

⁶ <https://portal.phenomenal-h2020.eu/home>

⁷ <http://phenomenal-h2020.eu/home/wp-content/uploads/2016/09/D6.1-final.pdf>

⁸ <http://phenomenal-h2020.eu/home/wp-content/uploads/2016/09/D6.2-PhenoMeNal-Virtual-Research-Community-Gateway.pdf>

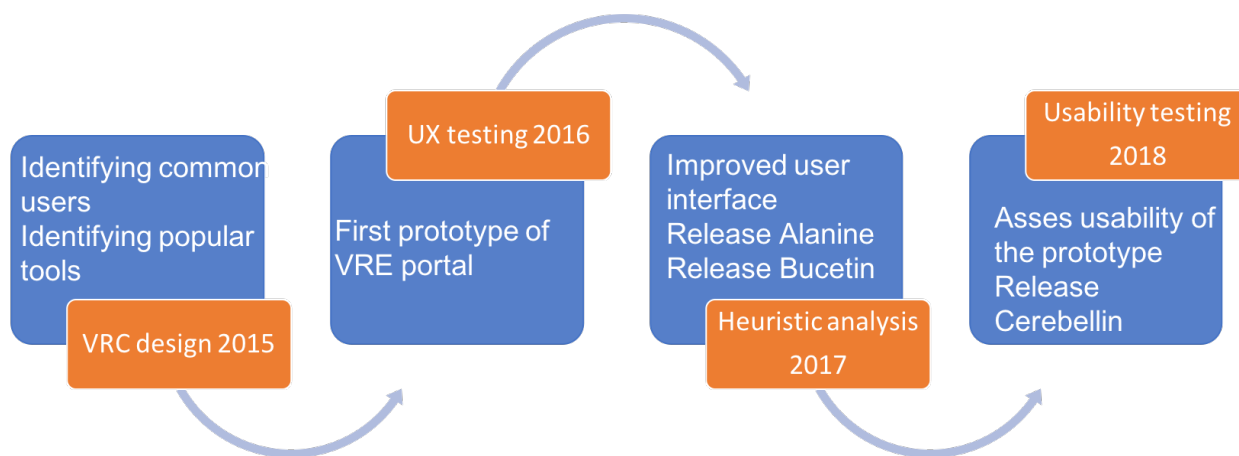


Figure 3. Usability testing process for PhenoMeNal VRE. The tests were carried out by a professional UX expert and the changes were implemented as part of the PhenoMeNal release cycle⁹ involving participation from all partners.

3.1.2.2 PhenoMeNal working groups

The activities of the PhenoMeNal working groups¹⁰ have been, and are, instrumental to provide connections between “core” metabolomics research and other communities. The overall aim is to facilitate the adoption of PhenoMeNal as a platform for metabolomics data analysis for scientists who are less familiar with complex IT systems, e.g. required to configure a suitable computing environment. Within this context, the activity of the working groups is also beneficial to exemplify and make the case for using metabolomics data within the routine work of other scientific fields.

The working group on **systems biology** includes the groups of Vitor Martins dos Santos (Wageningen University)¹¹ and Marta Cascante (University of Barcelona) as members of the ISBE consortium; the latter group is also a member of the PhenoMeNal consortium. A joint publication “*From correlation to causation: analysis of metabolomics data using systems biology approaches*”¹² has been accepted in journal *Metabolomics* in April 2018.

⁹ <https://github.com/phnmnl/phenomenal-h2020/wiki/Release-information>

¹⁰ <http://phenomenal-h2020.eu/home/about/management-structure/work-groups/>

¹¹ <https://www.wur.nl/en/Persons/Vitor-Martins-dos-Santoss.htm>

¹² <https://link.springer.com/article/10.1007%2Fs11306-018-1335-y>



The **clinical working group** includes the PhenoMeNal partners, the Hospital Clinic of Barcelona¹³ as well as the Departments of Medicine or Biomedicine from different institutions in Europe (Uppsala University, Sweden; University of Florence, Italy; CIBER de Enfermedades hepáticas y digestivas, Madrid; Imperial College London). The group was established to contribute views and suggestions on the kind and/or specifics of computational services needed to support the transition of metabolomics from clinical research towards clinical practice. The members of the group work together to identify the needs of the clinicians from metabolomics point of view and advise on functions required in PhenoMeNal tools and workflows by clinical users.

In practice, the contribution of the working groups to the sustainability of PhenoMeNal extends beyond the lifetime of the current project, both by providing inputs to the development of the overall platform to meet potential needs from other communities, and by showcasing applications of metabolomics that PhenoMeNal supports to highlight their usefulness in contexts different from ordinary metabolomics research. In particular, the clinical working group has provided an extensive vision of the road towards broader adoption of metabolomics within clinical settings, and of the support by PhenoMeNal to that vision.

3.1.2.3 PhenoMeNal Training program

The PhenoMeNal training strategy is firmly focussed on actively supporting and training users to ensure that the PhenoMeNal infrastructure is accepted and used by the community. Development of a successful and sustainable training strategy calls for more than just standardised high quality training materials. It requires appropriate guidance for trainers, as well as a framework to ensure its development, deployment, and administration. A well-structured training strategy should also ensure continuous availability of training for potential users. As such, the PhenoMeNal training strategy incorporates four components as shown in figure 4. These training programs are advertised through blog posts, social media and direct involvement with the national and international metabolomics communities: Metabolomics society¹⁴, Cambridge metabolic network¹⁵, London metabolomics network¹⁶, Scottish metabolomics network, EGI community¹⁷, French metabolomics community (RFMF) and ELIXIR metabolomics community.

¹³ <https://www.hospitalclinic.org/en>

¹⁴ <http://metabolomicssociety.org/events/upcoming-conferences-workshops>

¹⁵ <http://www.metabolism.cam.ac.uk>

¹⁶ <http://londonmetabolomicsnetwork.com/phenomenal-gateway-portal-to-metabolomics-data-analysis-in-the-cloud>

¹⁷ <https://www.egi.eu/blog/workshop-the-phenomenal-gateway-a-portal-to-metabolomics-data-analysis-in-the-cloud/>



In order to prevent redundant work, links to already existing and relevant training materials that fit within the scope of PhenoMeNal are also provided e.g coordinations with the Galaxy Training Network¹⁸ and the associated tutorials¹⁹.

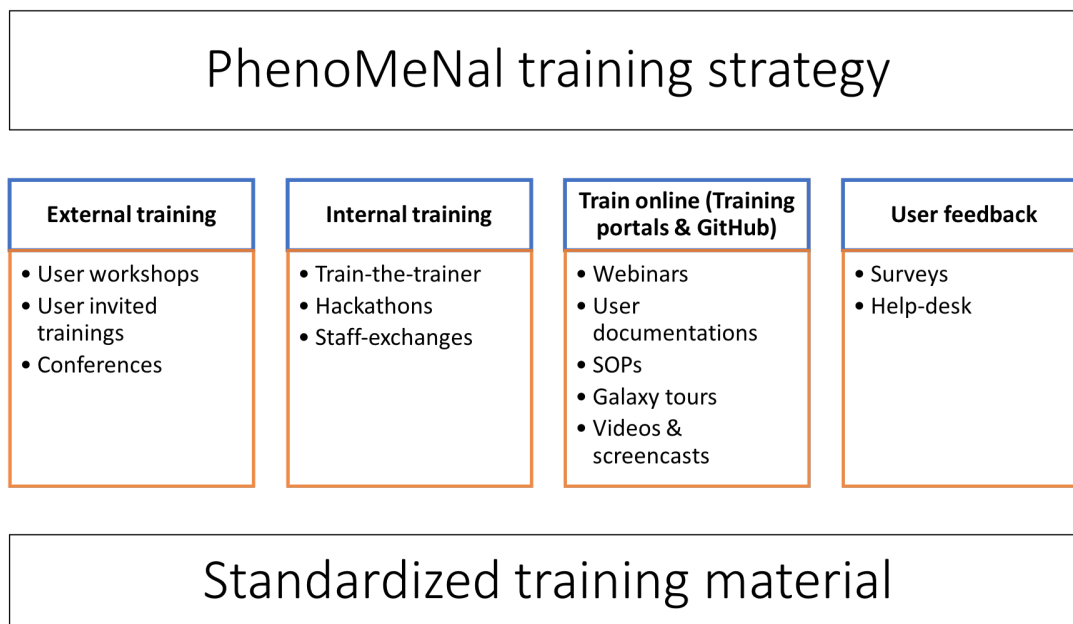


Figure 4. The PhenoMeNal training program is an all consortium effort and is intended for both infrastructure users as well as developers.

External training

External training includes face-to-face demonstrations, hands-on training, and seminars. These are delivered through user workshops²⁰, user invited training sessions and talks/workshops at conferences²¹. User surveys were conducted post workshop to gather user feedback²². For a detailed report on user workshops and feedback gathered see D1.4.5 (Bi-annual progress report)²³.

¹⁸ <https://galaxyproject.org/teach/gtn/>

¹⁹ <http://galaxyproject.github.io/training-material>

²⁰ <http://phenomenal-h2020.eu/home/about/events/>

²¹ <http://phenomenal-h2020.eu/home/about/presentations-posters/>

²² <http://phenomenal-h2020.eu/home/2017/10/31/cloudmet-2017-helps-establish-phenomenal-as-an-effective-data-analysis-and-training-platform/>

²³ <http://phenomenal-h2020.eu/home/wp-content/uploads/2018/03/Deliverable-1.4.5.pdf>



31
OCT

CloudMET 2017 helps establish PhenoMeNal as an effective data analysis and training platform

By Namrata No Comments Communications, General, Workshops All, Outreach, Surveys

It's been a bit more than a month since the *Scientific School on Cloud-based Metabolomics Data Analysis and Collaboration* (CloudMET) 2017 wrapped up. This event distinguished itself from the various metabolomics data analysis schools as its organizers – CRS4 and the rest of the PhenoMeNal consortium, with great help from the University of Cagliari – designed a training program that merged state-of-the-art data analysis techniques with notions about how to put these into practice with cloud computing technologies. The PhenoMeNal platform was central to this latter point, as it makes cloud computing infrastructure more easily accessible to everyone, and it provided the foundation on which many of the practical sessions of the course were built.

Figure 5. Feedback gathered at CloudMET 2017 helped to assess the impact of the project. The organisation of the workshop also offered the first stress test of the PhenoMeNal application stack²⁴.

PhenoMeNal has also been advertised across diverse platforms, e.g. the recently established Central German Galaxy User Workshop in April 2018 at IPB.

Internal training

PhenoMeNal is a large consortium effort involving a number of partners each with their own responsibilities. As a result, individual members/groups do not necessarily have detailed knowledge about all the broad range of PhenoMeNal aspects and features. Our internal training strategy provides opportunities for consortium members to learn about and familiarise themselves with all aspects of the PhenoMeNal project. Internal training is provided through train-the-trainer sessions, hackathons and staff exchanges. The sessions are recorded and made available on PhenoMeNal YouTube²⁵ and Google site²⁶ (for internal usage and comments).

- Train-the-trainer sessions are aimed at all consortium members and involve presentations/demonstrations about individual PhenoMeNal aspects such as; the

²⁴ <http://phenomenal-h2020.eu/home/2017/10/17/first-stress-test-of-the-phenomenal-application-stack-running-on-the-de-nbi-cloud/>

²⁵ <https://www.youtube.com/channel/UCXGAvsVNQk-aUpckjRC8Ang>

²⁶ <https://sites.google.com/view/phenomenal-internal-training/>



PhenoMeNal portal, setting up the infrastructure, data-processing tools/workflows and statistical tools/workflows. The aim of these sessions is to give consortium members a detailed and solid overview of the various aspects of the PhenoMeNal infrastructure so that when potential PhenoMeNal users/interested parties are engaged, all aspects can be explained adequately.

- Hackathons are aimed at groups of consortium members who are involved in similar research and development activities e.g. mass spectrometry and/or statistics. The primary aim is the continued development of the PhenoMeNal infrastructure, but they also allow for participants to gain detailed knowledge about PhenoMeNal features directly from the developers. In a similar way, staff exchanges allow for transfer of knowledge between consortium members.

Online training materials

A number of channels for online training are provided, which allow for users to learn about PhenoMeNal.

- The PhenoMeNal portal webpage²⁷ includes detailed user documentation on all aspects of PhenoMeNal. The documentation is based on our GitHub Wiki²⁸ that can be easily updated and developed further beyond the timescale of the project.
- A series of webinars²⁹, each describing a specific PhenoMeNal aspect, tool or workflow in detail have been delivered by consortium members. Webinar recordings are also available on the EMBL-EBI training website³⁰ and ELIXIR Tess portal³¹.
- Galaxy tours³², accessed via the PhenoMeNal portal, provide interactive step-by-step online guidance about Galaxy tools or workflows. Tours can easily be edited when tools or workflows get updated. Additionally, a number of videos and screencasts are also available via the PhenoMeNal and EMBL-EBI YouTube pages. These videos include demonstrations of PhenoMeNal features, tools and workflows.

Help Desk

²⁷ <https://portal.phenomenal-h2020.eu/help>

²⁸ <https://github.com/phnmnl/phenomenal-h2020/wiki>

²⁹ <http://phenomenal-h2020.eu/home/about/events/>

³⁰ <https://www.ebi.ac.uk/training>

³¹ <https://tess.elixir-europe.org>

³² <https://public.phenomenal-h2020.eu/tours>



The current helpdesk³³ will continue its current operation after the end of the current project period. However, it is expected that the helpdesk backoffice membership will vary due to various personnel commitments from individual partners.

Issues raised on GitHub related to a specific tool or functionality, will continue to be addressed in the same manner as today. The project assumes that the current maintainers of a respective GitHub repository will continue to support and develop their open source tools on a best-effort basis.

3.1.3 Stakeholder engagement

The PhenoMeNal stakeholder panel comprises of the representatives from the European Grid Infrastructure (EGI), other infrastructure initiatives, vendors, SME's and publishers. As part of its stakeholder engagement plan, the consortium profiled the stakeholder needs, interests and expectations from the project and has continuously worked towards collaborations for sustainability^{34,35}. Periodic stakeholder consultations has ensured that the services offered meet the requirements of the various user communities.

As part of its dissemination and outreach strategy, the consortium has also taken efforts to raise awareness for standards, services and tools by working closely with publishers to ensure that standards established and linked to services in PhenoMeNal can be readily adopted for publications³⁶.

3.1.3.1 European Open Science Cloud (EOSC) and GO FAIR initiative

As part of its objective to coordinate with national and international initiatives for sustainability, the consortium has worked towards establishing close collaborations with the EOSC. A 2-day strategic workshop entitled “Establishing a node for FAIRifying metabolomics and phenomics data in the European Open Science Cloud” was hosted by Leiden University (LU) in March 2017³⁷. As a successful outcome of the meeting, the “Metabolomics Implementation Network” of the European Open Science Cloud was launched. The network aims to implement the FAIR data principles in the field of metabolomics in Europe and is comprised of key members in the European metabolomics community, including EMBL-EBI, Imperial College London, Leiden University, Oxford University, IPB-Halle, CIRMMMP, CNRS, University of Jena, companies like Janssen

³³ <http://phenomenal-h2020.eu/home/help/>

³⁴ <http://phenomenal-h2020.eu/home/wp-content/uploads/2016/03/Revised-Deliverable-3.1.1.pdf>

³⁵ <http://phenomenal-h2020.eu/home/wp-content/uploads/2016/03/Deliverable-3.1.2.pdf>

³⁶ <http://phenomenal-h2020.eu/home/wp-content/uploads/2016/03/Deliverable-3.2-revised.pdf>

³⁷ <http://phenomenal-h2020.eu/home/about/events/fairifying-metabolomics-and-phenomics-data-in-the-eosc/>



Pharmaceuticals, DSM, the Hyve, and linked Data experts. A “Leiden declaration on FAIR metabolomics” was signed with the following objectives:

- To form and develop the open Metabolomics Implementation Network on metabolomics data and services; we invite others to join;
- To collectively implement standards compliant with FAIR principles in the wider research community and to actively communicate these;
- To work closely together with other communities on better capturing and understanding phenotypes enabling integrated approaches;
- To support the GO FAIR initiative with the widespread implementation of the FAIR principles in the EOSC, contributing to a global open internet of FAIR data and services.

On January 25th 2018, we participated in a meeting organised by the GO FAIR office, with other GO FAIR Implementation Networks, on Rare Diseases, Training, Personal Health Train and Biodiversity. In addition, we are seeking synergies with other “omics” communities (proteomics, genomics, bioinformatics), for joint actions to make “omics” data more FAIR at the source (where they are being generated) and more easily interoperable, to better answer biological questions in for instance Rare and Chronic Diseases. We will align the activities of the GO FAIR Metabolomics Implementation Network with those of the ELIXIR Metabolomics community, described in the next section.

3.1.3.2 ELIXIR metabolomics community

As part of its sustainability measures, the consortium has worked closely with the ELIXIR infrastructure. A *letter of support* from ELIXIR director in January 2017 express support for the PhenoMeNal project.

Following a joint workshop with ELIXIR “The Future of Metabolomics in ELIXIR” in Frankfurt in April 2017³⁸, and as a successful outcome, a white paper³⁹ was published in F1000. Subsequently, ELIXIR invited the community to draft an Implementation Study, to foster the interaction of the Metabolomics Community with the 5 existing platforms in ELIXIR (Data, Tools, Interoperability, Compute and Training) and thereby preparing for inclusion of the Metabolomics Community in the ELIXIR Work programme 2019-2023⁴⁰. This Implementation Study (budget EUR 140.000, with additional in-kind contributions from the partners of EUR 200.000 for an 18 month period) was approved in March 2018

³⁸ <http://phenomenal-h2020.eu/home/about/events/the-future-of-metabolomics-in-elixir/>

³⁹ <https://f1000research.com/articles/6-1649/v1>

⁴⁰ <https://www.elixir-europe.org/news/elixir-establish-new-use-cases-proteomics-metabolomics-and-galaxy>



by ELIXIR Heads-of-nodes. A kick-off meeting for the ELIXIR Implementation Study has been scheduled from 18-19th September 2018 in Rotterdam.

Several members of the PhenoMeNal consortium are also part of their national ELIXIR node:

EMBL-EBI

The public PhenoMeNal portal is hosted on the EMBL-EBI EMBASSY cloud infrastructure. EMBL-EBI will ensure the current allocated hardware and software infrastructure is available for at least two years after the end of the project (August 2018). EMBL-EBI also hosts the MetaboLights repository⁴¹, which is an ELIXIR recommended deposition database⁴² for metabolomics data. The PhenoMeNal VRE Portal is portable in the sense that the infrastructure can be, and currently is, deployed in its entirety on other geographic sites on various cloud technologies. Secure data transfer between PhenoMeNal VREs and MetaboLights is enabled independently of geographic location.

ELIXIR-Spain

Marta Cascante's team at UB is currently the official metabolomics unit inside the Bioinformatics Platform of the Instituto de Salud Carlos III (ISCIII), the Spanish node of ELIXIR. As a part of ELIXIR-Spain UB will receive support to deploy the PhenoMeNal tools in order to make them available through the ELIXIR portal. Barcelona will also continue to work on the maintenance of the existing Fluxomics workflow using resources as part of the ELIXIR-Spanish node.

ELIXIR-UK

The UK partners, UOXF and UoB, are represented under ELIXIR as national nodes. UOXF provides an important chunk of such presence as both ISA project, which drives the study metadata exchange under PhenoMeNal, and FAIRsharing project, which plays a central role in supporting assessment of FAIR datasets. In fact, FAIRsharing is currently involved in several IMI2 projects. Furthermore, the group is co-leading work on implementation metrics of the FAIR principle in both the EU and US (NIH DCCPC program). The ISA project under ELIXIR, has been a point of contact for the plant community and, through continued engagement with that community, it can provide a natural entry point for PhenoMeNal infrastructure and all the benefit of the work carried out under that project.

⁴¹ <https://www.ebi.ac.uk/metabolights/>

⁴² <https://www.elixir-europe.org/about-us/who-we-are/nodes/embl-ebi>



ELIXIR- Germany

As an initiative from IPB Halle, a German instance of the PhenoMeNal infrastructure runs in the de.NBI cloud. The de.NBI cloud⁴³ is a fully academic cloud, where academic cloud centers provide storage and computing resources for locally stored data. After completion, it will comprise more than 15,000 compute cores and 5 PB of storage capacity, provided by five local cloud centers in universities of Bielefeld, Freiburg, Gießen, Heidelberg and Tübingen, in close cooperation with the ELIXIR cloud. The team at IPB has deployed installations of PhenoMeNal on both the cloud center in Giessen and in Bielefeld. The PhenoMeNal instances on the de.NBI cloud were used in both the CloudMET 2017 and the “Workflows in Metabolomics” workshops. Other German groups, for instance of Christoph Steinbeck (U Jena) and Oliver Kohlbacher (U Heidelberg) are also involved and ELIXIR-Germany and are like to be involved in sustaining PhenoMeNal.

ELIXIR-NL

In ELIXIR-NL there is strong support for developing a national cross-omics Research Infrastructure (www.x-omics.nl), building on the experiences in PhenoMeNal. Leiden University is leading the Metabolomics activities, Utrecht University the Proteomics activities, UMC Utrecht the genomics activities and Radboud UMC the cross-omics and data-integration activities. Together these groups have successfully applied for a € 17 mln large-scale Research Infrastructure grant at the Netherlands Science Foundation, and work is conducted in close collaboration with the Dutch Techcentre for Life Sciences⁴⁴, which hosts the Dutch node of ELIXIR, ELIXIR-NL.

ELIXIR-SE

PhenoMeNal is represented in the Swedish ELIXIR node, which has lead to transfer of competence and the use of KubeNow for cloud contextualisation (developed in PhenoMeNal) to other ELIXIR-SE projects. ELIXIR-SE has indicated its strong support towards contributing to future PhenoMeNal efforts, including the matching funds for the ELIXIR Metabolomics case study. Furthermore, ELIXIR-SE is the member in the nordic Tryggve2 project⁴⁵ for working with sensitive data within the nordic countries where PhenoMeNal has been selected as one of the case studies during 2017-2020⁴⁶.

⁴³ <http://www.denbi.de/cloud>

⁴⁴ www.dtls.nl

⁴⁵ <https://neic.no/tryggve2/>

⁴⁶ <http://phenomenal-h2020.eu/home/2017/11/14/phenomenal-in-nordic-clouds/>



3.1.3.3 PhenoMeNal Industry Board

An industry workshop was hosted in June 2016 and tasked to engage instrument vendors to raise awareness of the PhenoMeNal initiative and to ensure optimal interoperability of the PhenoMeNal infrastructure and instrument vendor's data formats and tools. Following the nomination from the participants, an Industry advisory board⁴⁷ was established. Since the formation, the board has been invited to participate in joint meetings with the scientific advisory board (SAB) and stakeholder meetings. PhenoMeNal has also been presented from time to time in various workshops hosted by the EMBL-EBI industry program⁴⁸. As an expression of interest, a *letter of intent* has been signed by Bruker.

3.2 Technical Sustainability

Technical assessment is a vital part in selecting and developing the sustainable key components in an infrastructure like PhenoMeNal. Since the inception of the proposal, PhenoMeNal has focussed on integrating well established tools and standardised data formats. The work done to improve the functionality, scalability and interoperability of these tools in workflows has been documented in detail in the VRE portal. Moreover, some of the workflows are illustrated as interactive Galaxy tours. Naturally, all tools and workflows are available as open source on GitHub. Technical sustainability is also ensured by three different types of testing – container testing⁴⁹, data testing⁵⁰ and integration testing⁵¹. These tests are automatically and continuously run by our Continuous Integration Framework, and thus ensure that any defect or incompatibility arising between tools is detected immediately, when it is easiest to resolve. All integrated tools and their dependencies are encapsulated in software containers (e.g. Docker). Even when the framework of the e-infrastructure changes, tools can be run as they are binary compatible.

3.2.1 Security

Based on the reviewers comments in the M18 periodic review security has been taken up as a serious issue by the consortium. Scientific research is rapidly moving towards cloud computing, especially in computationally intensive areas like physics and life sciences. Work traditionally carried out in High Performance Computer Systems (HPC), inside secure intranets, is now also increasingly migrating to decentralized data centers run by

⁴⁷ <http://phenomenal-h2020.eu/home/about/management-structure/>

⁴⁸ <https://www.ebi.ac.uk/industry>

⁴⁹ <http://phenomenal-h2020.eu/jenkins/view/%20A.-%20Container%20tools/>

⁵⁰ <http://phenomenal-h2020.eu/jenkins/view/%20B.-%20Container%20data%20tests/>

⁵¹ <http://phenomenal-h2020.eu/jenkins/view/%20C.-%20Integration%20tests/>



various cloud providers. As such, security concerns that were deemed less relevant due to running inside a private network are now brought to light. The consortium has met with Tryggve2 project leader and their WP leader for security, which resulted in a security review meeting on March 13th 2018⁵² covering all technical aspects for a systems perspective (taking into account ELSI aspects as well). Work is ongoing to ensure that we are both aware of and address all security concerns.

3.2.2 PhenoMeNal VRE portal

The PhenoMeNal VRE portal has been designed with a modular architecture that facilitates its maintenance, extension and significantly reduces the learning period for new developers that want to approach it. Indeed, the portal development process smoothly sustained the transition from a resigning to a new member of staff, without any delay in the delivery schedule. This same modularity makes it easy to make the portal compatible with new cloud computing providers. In fact, the consortium is currently adding support for PhenoMeNal deployments at two new nordic partners: the Swedish National Infrastructure for Computing (SNIC) and the Finnish IT Center For Science (CSC). Moreover, the VRE portal leverages modern software container technology to make it robust to future changes to the execution platform – e.g., software updates without backwards compatibility. As a result of its design decision, the portal will be executable and useful as long as software container platforms are available.

In addition, a significant amount of effort was spent on ensuring a positive user experience (UX) to further encourage long-term adoption by users. As described earlier in the document, the design of the portal has placed usability as a top priority since its inception and the work has profited from frequent evaluations from professional UX experts and formal UX testing.

A current focus point for the technical sustainability of the VRE portal is its dependence on EMBL-EBI's Technology & Science Integration (TSI) portal – currently a closed-source service run by EMBL-EBI's TSI group, which is used to implement parts of the PhenoMeNal VRE portal's underlying functionality. However, the consortium is currently addressing this issue, with the plan to package and open the TSI portal's source code in the coming months and thus allow the service to be run by any interested parties. The EMBL-EBI TSI group is part of the project through the overall EMBL-EBI coordination.

⁵² <http://phenomenal-h2020.eu/home/2018/02/13/risk-assessment-for-using-sensitive-data-in-the-cloud/>



3.2.3 Sustainability measures for (individual) tools and containers

Deliverable 9.4⁵³ provides an update on existing software tools, workflows and analytical pipelines supported in PhenoMeNal, covering the new developments since the *D9.1 Report on existing software tools, workflows and analytical pipelines initially supported in the PhenoMeNal grid*⁵⁴. In the annex of this report a full list of all tools is given, that are currently integrated in PhenoMeNal. The list includes their status in the continuous integration and for the upcoming release and availability in the PhenoMeNal portal.

This section describes the measures taken by the consortium partners for the sustainability of tools and containers integrated in PhenoMeNal beyond its 3-year funding period.

3.2.3.1 EMBL-EBI

EMBL-EBI has led the design and implementation of the the Galaxy-Kubernetes integration, which is one of the main components underpinning the tools running on Galaxy on PhenoMeNal. This key integration is composed of a (1) Kubernetes runner for Galaxy which allows Galaxy to send jobs to a Kubernetes cluster, (2) a slim Galaxy container which includes all the dependencies and setup necessary for the runner to be executed within Galaxy as a container (and send jobs from within), (3) additional setup on the community maintained docker-galaxy-stable container images to run PhenoMeNal on Kubernetes using them and (4) Helm charts, which allows the automation and packaging of the deployment process of PhenoMeNal (using 2 or 3) or other Galaxy instances on a Kubernetes cluster through a single command line call. All components listed above are available as open source software, and points 1, 3 and 4 have been contributed to the Galaxy community's main software repositories and these contributions have been maintained during the duration of the project under their supervision. It has been suggested by some collaborators that the Galaxy-Kubernetes integration could be the single largest contributions by PhenoMeNal to the general Bioinformatics community (in comparison to individual tools). Its sustainability is guaranteed by the fact that most of this contribution has been left in the ownership of the thriving Galaxy community on open source Git repositories, which also makes them easily accessible. A manuscript is in preparation to illustrate the use of this integration in the wider context of Bioinformatics cloud computing across omics.

⁵³ <http://phenomenal-h2020.eu/home/wp-content/uploads/2018/03/Deliverable-9.4.pdf>

⁵⁴ <http://phenomenal-h2020.eu/home/wp-content/uploads/2016/09/D9.1.pdf>



EMBL-EBI also developed Galaxy wrappers, tool containers and tools code for accessing MetaboLights data (MTBLS Downloader, which is currently being folded into the more feature rich solution mainly developed by CEA partners) and visualizing metadata (MetaboLights Factors Viz, which allows users to visualize the interactions between different factors and their values in terms of number of samples). In addition, EMBL-EBI has been a key participant in shaping the requirements for the many of the ISA Galaxy tools developed by the University of Oxford.

EMBL-EBI was the initiator of the PhenoMeNal Portal, its web application code, backend database for metadata, integration with Single Sign-On, integration with the EMBL-EBI Cloud Portal and with KubeNow deployment logic. All of the components of the portal, as in the Galaxy case, are containerised and deployable through Helm charts, which improve its usability across different cloud providers. Lately, these elements have been transferred successfully to CRS4.

For the work described above, sustainability is addressed by the fact that all of the tools, frameworks and wrappers are available through docker containers, which means that they can easily be deployed and executed regardless of hardware setups used for many years to come. EMBL-EBI is controlling the process for integrating with MetaboLights to ensure seamless dataflow from tools like the MTBLS Downloader and the ISA/MetaboLights Factor Viz. For the latter, this tool can also work with outputs available from other tools that start from an ISA archive, enabling more flexibility. Naturally further development of these tools will require efforts beyond the end of the current project period.

In terms of sustainability, EMBL-EBI has been key in liaising with the Biocontainers community, so that they can adopt all the PhenoMeNal tool containers as their own, and then can ensure their continuous development/maintenance through their community. For the past two official releases, the released versions of PhenoMeNal containers were deposited under the Biocontainers docker hub namespace and since the last release, all PhenoMeNal containers GitHub repos have been imported as submodules inside Biocontainers as well. EMBL-EBI has also shared the continuous integration (CI) setup deployed on the EMBL-EBI EMBASSY Cloud with Biocontainers administrators, so that they can replicate the setup to build both Biocontainers and PhenoMeNal containers automatically on their own OpenStack setup, improving the sustainability of PhenoMeNal tool containers building within the Biocontainers community.



3.2.3.2 IPB

IPB has been involved in the development of a variety of open source software tools used for Metabolomics analysis over several years now. The high demand within the community and the rapidly advancing field of application further strode ahead the enhancement and applicability of the software farm. With the integration of these tools into PhenoMeNal these tools reached a new form of applicability as an essential part of the e-infrastructure as they are used by several other tools and integrated into workflows that build upon them.

The IPB also manages several tool dependencies in PhenoMeNal, for example: rbase, XCMS 3.0 + 1.x and CAMERA that are used by many mass-spec related tools; rbase, speaq, nmrm1conv, nmrm1d for NMR related tools. Dependencies are important for our microservices architecture to keep containers small and efficient. Many of the tools listed below are also used in Galaxy workflows, e.g. NMR workflow (rbase, speaq, nmrm1conv, nmrm1d (NMRProcFlow)), MetFrag (xcms, camera, metfrag-cli, metfrag-cli-batch), and the workflow for “Metabolomics LCMS/MS processing, quantification, annotation, identification and statistics” (OpenMS, MSnbase). We are developing the workflows in parallel with W4M to ensure sustainability and to promote the tools and workflows to a broader audience. Within PhenoMeNal the IPB team promoted the usage of community-driven frameworks like BioContainers and now has started integrating them as basis for the PhenoMeNal software containers (like MetFrag).

With MetFrag we have also started to “flip” the container and Galaxy tool development around. At the beginning of PhenoMeNal, many of our underlying technologies and concepts were not yet widespread in the different communities. Hence, we created container descriptions ourselves, and created Galaxy tools based on these containers. We have now migrated the MetFrag dependency handling to a Bioconda package, from which in turn automatically Biocontainer images are created. What is still missing in Biocontainers are the testing and continuous integration strategies in PhenoMeNal. Thus, we still build a PhenoMeNal container on top of the Biocontainers which is then integrated into the PhenoMeNal release process. The MetFrag Galaxy tool description is now maintained as part of the Galaxy IUC community, and together with the Bioconda package allows to use MetFrag also on other Installations in the Galaxy and Metabolomics community.

Sustainability in terms of commercial cloud provider cost for computing and hosting has been relieved in the short term with PhenoMeNal being awarded \$20 000 1 year free research credits on AWS (Successful application from IPB/EMBL-EBI). Additionally ICL has secured free credits on Microsoft Azure, this is described under the ICL section later in this report.



3.2.3.3 UOXF

The FAIR Data Science team at the **Oxford e-Research Centre** has been developing the ISA framework⁵⁵ for managing life science experiment data over the last 10 years. Most recently the Python ISA API⁵⁶ has been developed as a platform that allows software developers to build tools relating to the ISA data formats. The ISA API is entirely open source and developed independently of the PhenoMeNal project. The Galaxy tools that have been developed during PhenoMeNal have all been designed around the ISA API as the platform for the supported features. Additionally, the ISA-related Galaxy tools are being developed and hosted independently from the core PhenoMeNal source code repositories within the ISA-tools/isatools-galaxy⁵⁷ Github repository so that the longevity of the tool support extends beyond the PhenoMeNal project. Within the `isatools-galaxy` repository, there are two separate branches of tools: (1) Under the `phenomenal` branch, we keep versions of the tools that are specifically released in the PhenoMeNal software releases, and (2) under the main trunk of development in `master` and `develop` branches, we keep versions of the same sets of tools that are configured for running within normal Galaxy instances. By developing versions of the ISA-related PhenoMeNal tools to work in normal Galaxy, we will ensure that the lifetime of support of the tools is maximised, and we intend to publish these tools to the main public Galaxy toolshed⁵⁸ hosted by the Center for Computational Biology and Bioinformatics at Penn State University, USA.

Furthermore, by closely linking Galaxy-related work to data deposition to EMBL Metabolights, which also builds, in part, on the ISA-API developed under PhenoMeNal, we ensure the perenity of datasets and services FAIR data management. Indeed, the metadata tracking affording in Galaxy brings two critical components to data FAIRification, the process of making data Findable, Accessible, Identifiable and Reusable. First, it ensure datasets are structured around an accepted community syntax (ISA). Second, the Galaxy components have been enhanced to allow for advanced semantic markup based on open ontologies (the list of which can be found here⁵⁹) integrated under the Metabolomics Standardization Initiatives Ontology (MSIO⁶⁰) and use of compact URI referencing.

Work with UK National Phenome Centre has delivered export functions in ISA format from local databases as well as alignment of naming conventions. This further builds the

⁵⁵ <http://isa-tools.org/>

⁵⁶ <https://github.com/ISA-tools/isa-api/>

⁵⁷ <https://github.com/ISA-tools/isatools-galaxy>

⁵⁸ <https://toolshed.g2.bx.psu.edu/>

⁵⁹ <https://docs.google.com/document/d/1hFhdcEBtQ6fnfRZjl3QYPxwzhUWz88jtFfrLtcciwBE/edit>

⁶⁰ <https://github.com/ISA-tools/MSIO/>



technical sustainability of the network by ensuring potential for interoperation beyond the lifetime of PhenoMeNal project.

3.2.3.4 UU

KubeNow VRE deployment software

KubeNow is designed to have a minimal and easy to maintain own codebase and is instead integrating programs from large well maintained open source projects such as Terraform, Ansible and Kubernetes. In an effort to further improve the sustainability, KubeNow has been developed with a plugin system making it generically supporting other projects and therefore gaining a larger user community that is helping maintaining the project. The KubeNow project has so far been adopted by NBIS (National Bioinformatics Infrastructure Sweden), th Finnish CSC (IT Center for science) and within the The Nordic e-Infrastructure Collaboration (NeIC⁶¹) and in particular the Glenna2⁶² project. The PhenoMeNal plugin repository contains a very small codebase that is easy to maintain outside the generic KubeNow codebase. The CI-testing of the OpenStack components have been guaranteed until Feb 2019 via SSC (SNIC Science Cloud). The CI-testing on Amazon, Google and Microsoft Azure Cloud providers are not guaranteed after the project. KubeNow is adopted as a core component and will be further developed within the H2020 OpenRiskNet project (2017-2020).

3.2.3.5 UB

Several tools implemented as computational support for fluxomic workflow aimed at the evaluation of metabolic fluxes in cells, which are consistent with the distribution of ¹³C isotopes in various intracellular metabolites from substrates artificially enriched with this stable isotope. There are three subsequent steps in this workflow.

i) Extraction of mass isotopomer distribution (MID) of the metabolites of interests from the time course of m/z data registration, saved by mass spectrometers in raw NetCDF files. In addition to Ramid, implemented for the first release, **Cdf2mid** functions for Cerebellin. It designed for new data, not presented in MetaboLights. It reads more complex than Ramid NetCDF files containing several simultaneously registered m/z spectra of metabolights, in addition to the NetCDF files, it uses a simple, easy to describe information about the experimental conditions, and provide the output MID, easily aggregable in MetaboLights. Moreover, two new tools are prepared for the next release: **Simid**, which reads the most complex NetCDF, containing the spectra simultaneously registered for all metabolites of interest in SIM (selected ions monitoring) mode, and

⁶¹ <https://neic.no>

⁶² <https://wiki.neic.no/wiki/Glenna2>



Scamid, which reads the NetCDF, containing the spectra for all metabolites of interest registered in scan mode. This set of tools fulfills the necessity of MID extraction from the possible types of NetCDF files.

ii) Correction the raw MID for the natural abundance of ^{13}C and other isotopes affecting the measured values, and also a possible overlapping of m/z peaks for various metabolites⁶³. The tool **Midcor** made compatible with the new tools for MID extraction, described above.

iii) Fitting the corrected MID using steady state or dynamic simulation of the metabolic fluxes, thus evaluating the set of metabolic fluxes consistent with the MID. An improved version of **Isodyn**, a tool for dynamic simulation of corrected MID, is prepared for the next release. Its extended functionality allows changing the scheme of an underlying kinetic model, which is a base for MID simulations, and account for the change of total metabolite concentration. It is also compatible with the old and new tools for MID extraction and correction, and it will substitute the previously implemented version.

Regarding the workflow that is already running, UB has improved the velocity of isotoflux and are preparing this tool for the analysis of the models provided by Molecular Connections.

3.2.3.6 CEA

The 8 tools brought by **CEA** in the PhenoMeNal project for post-processing (signal drift and batch effect correction, normalization, transformation, quality control) and statistical analysis (univariate hypothesis testing, multivariate analysis, feature selection) have already been part of the Workflow4Metabolomics platform (W4M) for several years now. A team of developers and scientifics from INRA, CEA and CNRS are responsible for the maintenance of those tools, provide a help desk, and give annual “Bring Your Own Data” courses to train experimenters since 2015. Conversely, their integration into PhenoMeNal further increase their technical sustainability (thanks to the containerization and continuous integration features) and their use by the experimenter community (e.g., statistical trainings during the PhenoMeNal workshops and webinars). Furthermore, two new tools have been developed by CEA during the project, to connect the MetaboLights reference data repository and the Galaxy workflows. More specifically, a MetaboLights downloader and a converter from ISA to W4M formats have been integrated into PhenoMeNal: studies from MetaboLights can be now directly imported into PhenoMeNal Galaxy workflows (e.g., to perform statistical analyses), on an interactive or batch modes. This work, which involved three partners (CEA, EMBL-EBI, UOXF) was made possible by the PhenoMeNal gathering of expertises in Galaxy workflows, MetaboLights database, and ISA formats. It is a breakthrough for the sustainability of the network of

⁶³ BMC Bioinformatics, 2017, 18, 88



complementary e-resources, since all teams can now join their forces to provide an interoperable data analysis service to the community.

3.2.3.7 INRA

INRA contribution on tools was dedicated to the analysis of metabolomics data (identified metabolites) in the context of genome scale metabolic networks. One of the main challenge tackled during the course of the project was to map data in the networks since metabolite identifiers used in the networks don't necessarily fit the ones used in metabolomics datasets. A specific function was developed in a dedicated container to perform identifier searches. A second container was then developed to allow performing mapping on genome scale metabolic networks provided as SMBL file (for now principally works on Human since it was the first scope of the project). This container is implemented in the PhenoMeNal Galaxy instance. Finally a series of webservice had been developed to access MetExplore⁶⁴ database and functions.

3.2.3.8 Leiden University

(internal) Use case of PhenoMeNal after the end of the project

The BioMedical Metabolomics facility at Leiden University has been developing and running analytical platforms for over 10 years, measuring several classes of metabolites which together cover most of the human metabolome. A comprehensive list of the different platforms and the metabolite classes covered can be found at <http://bmfl.nl>⁶⁵. To support these platforms, a collection of tools named “mzquality”⁶⁶ have been developed to monitor the quality, to make corrections and report results derived from mass spectrometry measurements. Over time these tools became harder to manage and maintain as they were written by different people, using different software languages, and often impossible to use without expert help. With the help of (bio-)informatics and large-scale research infrastructure experts/partners within the PhenoMeNal consortium we have been able to define and adopt a tool development procedure⁶⁷ which helps us to develop and deliver the services required by the people in the lab in a sustainable manner. The first tools have been (re-)written in this way, are suited for running within the PhenoMeNal ecosystem, and will be included in upcoming release(s). For our facility this ensures the tools are versioned, properly tested, well documented, and peer reviewed. This way of developing tools will help us to focus on the problem at hand, and providing our employees with the means for delivering traceable and reproducible results. We are

⁶⁴ www.metexplore.fr

⁶⁵ <http://bmfl.nl>

⁶⁶ <http://mzquality.nl/>

⁶⁷ <https://github.com/phnmnl/phenomenal-h2020/wiki/How-to-make-your-software-tool-available-through-PhenoMeNal>



committed to keep supporting and improving the tool development procedure as described by PhenoMeNal as it is crucial for our facility to deliver fully validated results.

3.2.3.9 University of Birmingham

The University of Birmingham has contributed containers for three tools which did not have existing docker containers to the PhenoMeNal project; DIMSpy, mzml2isa and nmrml2isa. The integration of these tools into PhenoMeNal increases their sustainability and visibility. DIMSpy is widely used within the PhenoMeNal group in Birmingham, and a galaxy wrapper has been previously developed. The DIMSpy galaxy wrappers have been published to the main public Galaxy toolshed, further enhancing its sustainability beyond the lifetime of the PhenoMeNal project. The mzml2isa and nmrml2isa tools and their associated Galaxy wrappers are maintained in the ISA-tools Github repository as detailed in the UOXF section above.

Tool development procedure

UI for our tools

By following the procedure and guidelines as defined within PhenoMeNal it is much easier for us to provide an easy to use interface for our end users. With Galaxy being the default UI for running individual tools or complete workflows that consist out of a collection of tools that together form the basis of the pre-processing.

Infrastructure

The "Infrastructure as code" approach, to deliver on-demand installations of virtual research environments (VRE), has been adopted by our facility and institute. It allows more efficient resource sharing and, similar to the tool development procedure, provides versioning of the underlying research infrastructure used for data pre-processing and analysis. The code and templates being developed can be used to deploy a DRE to virtual machines, local workstation, or larger compute resources such as (HPC) cloud. Which again supports our ambition to be able to scale up to both commercial and national research compute resources (e.g. SURFsara, AWS, etc.) when local resources are not available or sufficient.

3.2.4 Data Formats and Terminologies

A number of syntactic and semantic resources have been developed or selected by PhenoMeNal. Those built or contributed to under PhenoMeNal (nmrML, nmrCV, mzTab-M) have addressed a coverage gap and are maintained by long term stakeholders. The remainder of the resources used by PhenoMeNal tools reuse existing efforts (ISA, mzML for the syntax and a collection of OBO Foundry ontologies, which behave like software



libraries for the semantic web, and are designed to interoperate). This deliberate strategy to reuse rather than reinvent is meant to directly address the issue of sustainability. The choices made to contribute to existing efforts, which have been self-sustaining for a number of years, ensure resilience and sustainability of the work while also securing the acknowledgement of PhenoMeNal's contributions to these efforts. See Annex 1 for tables that summarise the format and controlled terminologies implemented by the tools developed by our project.

3.2.5 Workflow engines

PhenoMeNal has adopted community-standard workflow engines for its e-infrastructure: Jupyter, Luigi and Galaxy. Jupyter and Luigi were adopted in their original form by running community-maintained Docker images with the PhenoMeNal e-infrastructure. On the other hand, the Galaxy platform has been extended by the consortium to provide proper integration with the Kubernetes infrastructure and support data formats that are important for the metabolomics community. These extensions to the platform have been contributed back to the original project and most of them have already been integrated. In fact, while PhenoMeNal in the past used to run custom-built Galaxy Docker images, as of the latest release the e-infrastructure is using community-built and maintained Docker images for the Galaxy workflow engine. By contributing its work to the community and by using community-maintained images, the PhenoMeNal project ensures that its CRE will be sustainable by the community at large after the project funding period comes to an end.

3.2.6 Deployment of PhenoMeNal VREs

3.2.6.1 EMBL-EBI EMBASSY Cloud

EMBL-EBI will ensure the current allocated hardware and software infrastructure is available for a time period of up to two years after the end of the project (August 2018). Support and maintenance of this OpenStack tenancy will be a collaborative effort between all the existing project partners.



3.2.6.2 EGI

Led by UU, the deployment of the PhenoMeNal VRE is now supported on EGI federated cloud resources. This opens up for users with access to EGI resources to deploy PhenoMeNal VRE. This will be an important achievement in light of the emerging EOSC and EOSC-Hub project where EGI has a leading role. A joint article for the EGI newsletter is currently under preparation.

3.2.6.3 EOSC-Hub

A collaboration with the EOSC-Hub project have been initiated with the intent of including the PhenoMeNal VRE as service maintained by the EOSC-Hub project to sustain and promote the tools to broader, disciplinary and interdisciplinary users. The PhenoMeNal project was presented by UU at the EOSC-Hub all hands meeting 16-17 April 2018, at the session “Leading e-infrastructure project communities into EOSC”.

3.2.6.4 OpenStack

The PhenoMeNal project applied successfully for resources on the german de.NBI OpenStack instances. The application was approved for 36 month, starting in May 2018.

3.2.6.5 AWS

The PhenoMeNal project applied successfully for cloud credits through the “Amazon Research Credits” scheme. In total, cloud credits equivalent to \$20,000 are available to be spend through PhenoMeNal as part of trainings, benchmarking, running a public instance and for compute-intense projects.

3.2.6.6 Scalability Testing: Local Instance at ICL and Azure Funding:

The ICL team has worked on testing the scaling of computational tools in PhenoMeNal and have run extensive experiments on different cluster sizes. Scalability tests have been run on 4 different systems: 1) high performance desktop, 2) Local 80 core server, 3) Microsoft Azure Cloud (~300 vCPUs), 4) EMBASSY Cloud at EBI (~1000 vCPUs).



As part of this work, the ICL team were successful in obtaining credits for the “Microsoft Azure for Research” programme, worth \$20,000 for 12 months. This enabled us to add a widely used commercial cloud platform to our scalability work, a very important aspect to demonstrate the ability of PhenoMeNal to deploy on standardised public cloud infrastructure.

Another part of the scalability testing work was run on a local instance of PhenoMeNal based at ICL. This is a local server with 80 vCPUs and 1Tb of RAM and is intended to be a resource for the analysis and processing of both public and patient sensitive data generated locally at ICL. The server is located behind ICL’s institutional firewall and is therefore a prime example of “bringing compute to the data” in order to address privacy and ethics considerations. This infrastructure is funded by the UK’s Medical Research Council as part of the UK Medical Bioinformatics Partnership Programme ([UK Med-Bio](#)) funded to March 2019, and is expected to be maintained beyond that through the UK Health Data Research initiative ([HDR UK](#)).

3.2.6.7 Other deployments

The PhenoMeNal VRE was deployed on the private cloud shared by CIRMMP and the Da Vinci European biobank in Florence. The private cloud is based on Mitaka Version of OPENSTACK, with the following resources: 64 core, 128 GB RAM 1.3 TB hard disk. In particular, Ubuntu 16.0.4 was installed on a instance of openstack and configured for the deployment of the service. The deployment was successfully achieved following the instructions given on GitHub⁶⁸.

4 Delivery and Schedule

The delivery is delayed: No

5 Conclusion

The partners in PhenoMeNal are committed towards the continuity and further development of the tools and services offered by PhenoMeNal after the end of the funding period and initiatives towards widening and deepening the use of PhenoMeNal tools and services. The partners have all voted to set up a foundation, as a sustainable

⁶⁸ <https://github.com/phnmnl/cloud-deploy-kubenow/blob/master/README.md#get-phenomenal-kubenow>.



administrative/management entity that will steer and control all aspects of the PhenoMeNal service portfolio and align these with other community initiatives (for instance the GO FAIR Implementation network and the ELIXIR Metabolomics Community). Establishing the foundation has been delayed (scheduled M30), because of interactions with legal departments of PhenoMeNal partners, which proved to be constructive yet time-consuming. In the technical sustainability part of this report an overview is given on the measures taken by the consortium partners for the sustainability of tools and containers integrated in PhenoMeNal beyond its 3-year funding period, as well as how the reviewers' recommendations on security have been implemented.



6 Annex 1

Data Exchange Formats	Description and Organisation
ISA formats	<p>ISA-Tab and ISA-JSON are the tabular and JSON-based formats that supports the management, exchange and submission of omics data. (ISA Commons community, led by UOXF)^[1]</p> <p>An RDF serialisation can be produced in 2 ways. Either by using the JSON-LD context file or by using the java code LinkedISA to convert ISA-Tab documents to RDF.</p>
MAF	<p>Metabolite Assignment File: a tab delimited format developed by EMBL-EBI Metabolights to support deposition of Metabolites identified over the course of an experiment. Based on mzTab 1.0, will update to the more metabolomics focused mzTab 1.1 when this is released by HUPO PSI</p>
mzML	<p>XML formats for encoding raw spectrometer output. PSI (Proteomics Standards Initiative)</p>
mz5	<p>HDF5 binary implementation of mzML, to reduce storage overhead in large-scale datasets.</p>
mzTab	<p>mzTab supports the reporting of technical/biological replicates within experimental designs using an adaptation of the system originally developed for mzQuantML. PSI</p>
nmrML	<p>NMR equivalent of mzML, mainly for metabolomics raw data and analytical chemistry of small molecules. COSMOS (Coordination Of Standards for MetabOlomicS)</p>



nmrTab	NMR equivalent of mzTab, mainly for metabolomics. <i>In specification</i> by COSMOS
SPREC code	a fingerprint coding human biological samples, isolation, storage and processing conditions
SBML	Systems Biology Markup Language. Standardise description of metabolic networks including model parameters.

Table 1.1: Non-exhaustive list of data exchange formats which will form the basis of the PhenoMeNal primary data exchange.

Controlled Terminologies	Description and Organisation (An ontology is a computational representation of a domain of knowledge based upon a controlled, standardised vocabulary for describing entities and the semantic relationships between them.)
NCBITaxonomy	The most complete organism taxonomy, co-maintained by NCBI and EMBL-EBI. Actively used and a requirement for annotation of any dataset in MetaboLights and other EMBL/ELIXIR databases.
UBERON	An integrated cross-species anatomy ontology covering animals and bridging multiple species-specific ontologies.
DOID	a comprehensive classification of human diseases organized by etiology
HPO	The Human Phenotype Ontology (HPO) aims to provide a standardised vocabulary of phenotypic abnormalities encountered in human disease. Each term in the HPO describes a phenotypic abnormality, such as atrial septal defect. The HPO is currently being developed using the medical literature, Orphanet, DECIPHER, and OMIM



CL	The Cell Ontology is a structured controlled vocabulary for cell types in animals.
CHEBI	Chemical Entities of Biological Interest. An ontology and database support the reporting of identifications of metabolites but also used to semantically markup any chemical compound used in an experiment
CHMO	CHMO, the chemical methods ontology, describes methods used to collect data in chemical experiments, such as mass spectrometry and electron microscopy prepare and separate material for further analysis, such as sample ionisation, chromatography, and electrophoresis synthesise materials, such as epitaxy and continuous vapour deposition It also describes the instruments used in these experiments, such as mass spectrometers and chromatography columns. It is intended to be complementary to the Ontology for Biomedical Investigations (OBI).
DUO	An ontology based on the consent codes description from the Dyke et al. paper at http://dx.doi.org/10.1371/journal.pgen.1005772 . It allows to semantically tag datasets with restriction about their usage, making them discoverable automatically based on the authorization level of users, or intended usage. This resource is based on the OBO Foundry principles, and its use is under review by the Broad Institute and the European Genome-phenome Archive (EGA) at EMBL-EBI
OBI	The Ontology for Biomedical Investigations (OBI) project is developing an integrated ontology for the description of life-science and clinical investigations.
PSI-MS	A structured controlled vocabulary for the annotation of experiments concerned with proteomics mass spectrometry. Developed by the HUPO Proteomics Standards Initiative (PSI).



nmrMLCV	<p>nmrML is an open mark-up language for NMR raw and spectral data. It has recently seen its first release, ready for public use. The development of this standard was previously coordinated by the COSMOS FP7- Project. It is now maintained within the PhenoMeNal H2020 project, setting up metabolomics data analysis e-infrastructures. The nmrML data standard is approved by the Metabolomics Standards Initiative and was derived from an earlier NMR XML format that was developed by the Metabolomics Innovation Centre (TMIC). It follows design principles of its role model, the mzML data standard, created by HUPO-PSI community for mass spectrometry.</p>
STATO	<p>STATO is a general-purpose STATistics Ontology. Its aim is to provide coverage for processes such as statistical tests, their conditions of application, and information needed or resulting from statistical methods, such as probability distributions, variables, spread and variation metrics. STATO also covers aspects of experimental design and description of plots and graphical representations commonly used to provide visual cues of data distribution or layout and to assist review of the results.</p>

Table 1.2: A list of interoperable semantic resources organised by domain.