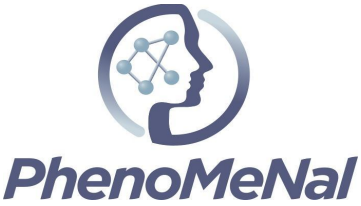

Deliverable 3.2

Project ID	654241
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	3
Work Package Title	Dissemination and Outreach
Deliverable Title	D3.2 Report on establishing and maintaining relations with publishers for supporting data deposition
Delivery Date	M16
Version	Revised
Work Package leader	UoB
Contributing Partners	UoB, UL, EMBL-EBI, UOXF
Authors	Ulrich Guenther, Michelle Thompson, Karen Atkins, David Johnson, Susanna Sansone, Philippe Rocca-Serra, Alejandra Gonzalez-Beltran, Peter McQuilton, Ralf Weber, Mark Viant, James Bradbury, Daniel Schober, Namrata Kale

Abstract:
This deliverable provides a report on journal policies and the consortium's strategy of establishing contacts with the publishers for supporting data submissions.

History of Changes:

- Evidence the progress made with the Public Library Of Science (PLOS)

editorial office.

- Evidence of progress in disseminating open data policies with a commercial software and service provider, Biocrates.
- Modified work plan

CONTENTS

1 EXECUTIVE SUMMARY

Dissemination in PhenoMeNal aims to establish close links between the consortium and the wider metabolomics, genomics and the biomedical communities in order to establish a large user base for the services offered. The goal is to raise community awareness and interests for the services provided by the PhenoMeNal consortium, from data and information mining, processing and analysis, but also for the use of omics technologies in a biomedical context, close to the clinic.

It is important that researchers in the field of metabolomics and related -omics develop awareness of, and interest to use, the tools and resources provided by PhenoMeNal. By fostering the incorporation of PhenoMeNal in the daily working practice of the users, we will achieve an effective adoption of eScience in the biomedical research community in addition to stimulating the adoption of best computational practices.

As part of the dissemination activities, it is essential that we work with publishers to ensure that standards established and linked to services in PhenoMeNal can be readily adopted for publications. While several journals have already subscribed to an open data policy, they need to be kept informed of new services in the field of metabolomics that actually facilitate the implementation of open data policies. In addition, some journals have yet to implement open data policies in metabolomics, and our outreach to such publishers is particularly crucial.

The reliance on EMBL-EBI MetaboLights and ISA formats, both endorsed by Springer Nature's *Scientific Data* and Oxford University Press/BGI *GigaScience* journals as “the” core repositories and representation formats, respectively, provides us with an excellent working model for high impact data publication and dissemination. Following this initial endorsement, via the Oxford BioSharing team, we have outreached to several other journals to promote the use of MetaboLights (see https://biosharing.org/biodbcore-000168). Here we describe these activities and in particular, evidence the progress made with the Public Library Of Science (PLOS) editorial office. We raise awareness of metabolomics relevant assay raw data standards, i.e. mzML and nmrML incl. controlled vocabulary developments, as sanctioned by the respective standardisation governance bodies, the Proteomics Standards Initiative (PSI) and the Metabolomics Standards initiative (MSI) respectively. The nmrML effort has recently been completed and outreach activities have started. In addition, we evidence our progress in disseminating open data policies with a commercial software and service provider, Biocrates.

2 WORK TOWARDS PROJECT OBJECTIVES

Activities described in this report contributes towards the following project objectives:

Objective 3.2: "Raise awareness for standards, services and tools provided by the PhenoMeNal grid".

3 DETAILED REPORT OF THE DELIVERABLE

We have focused our activity on high profile publishers that are seen as leaders in their respective field. Our philosophy is that when we are successful in bringing new tools and the PhenoMeNal e-infrastructure to their attention then a cascade of this awareness will follow across other publishers. In this section, we describe our progress with the Public Library Of Science[1] publishers, specifically as part of a collaboration with an industry partner, as well as with Springer Nature's Scientific Data[2] journal and GigaScience[3] publishers.

3.1 Public Library Of Science (PLOS) and Biocrates AG

Biocrates AG[4] is a leading provider of kits for targeted metabolomics in human plasma and urine. These kits are now routinely used for the rapid, quantitative assessment of about 200 metabolites with potential biomarker function[5]. Biocrates AG analytical products are complemented by an advanced software suite, Met*ID*Q[6], which offers an array of capability ranging from LIMS like functions, such as plate definition, data acquisition and logging to signal processing and data analysis allowing to perform univariate analysis or more advanced analysis such as PCA. We allowed to standardize Met*ID*Q outputs by delivering a ISA[7] export function through the ISA API (see below).
Our recent progress builds on the collaboration between Biocrates AG, EMBL-EBI resources (MetaboLights[8] and Chemical Entities of Biological Interest, ChEBI[9]) and the

[1] https://www.plos.org
[2] https://www.nature.com/sdata/
[3] https://academic.oup.com/gigascience
[4] http://www.biocrates.com
[5] http://www.biocrates.com/products/research-products/absoluteidq-p180-kit
[6] http://www.biocrates.com/products/software
[7] http://isa-tools.org
[8] http://www.ebi.ac.uk/metabolights/
[9] https://www.ebi.ac.uk/chebi/

University of Oxford, initiated under the EU FP7 COSMOS[10] project. Efforts by the PhenoMeNal developers to firmly establish a deposition workflow for targeted metabolic profiling datasets (generated using the kits manufactured by Biocrates AG) have resulted in the creation of several key deposition services:

- For two Biocrates kits (i.e. Absolute*ID*Q p150 and Absolute*ID*Q p180), the list of targeted metabolites has been submitted to ChEBI and re-annotated by the resource's curators to ensure chemical identities and unified annotation. A newly released Biocrates kit (June 2017), named Absolute*ID*Q p400[11], which measures more than double the number of metabolites and also probes lipid metabolism, is currently being processed following the same procedure. The integration of those metabolite lists ensures full compatibility and interoperability with ChEBI, EMBL-EBI nomenclatures and MetaboLights.
- The Investigation/Study/Assay-Application Programming Interface (ISA-API), a python library aimed at providing support to ISA format (i.e. the syntax used by EMBL-EBI Metabolights to archive study description) now features a service for dealing with Biocrates Met*ID*Q software outputs and producing syntactically valid ISA documents[12].

  Paired with updated ISA configurations and ISA-API validation services, submitting Biocrates generated datasets in the standardized ISA format is now supported more comprehensively, easing deposition of Biocrates data sets into MetaboLights.
- A collaboration with the PLOS editorial office has established EMBL-EBI MetaboLights as the recommended repository for enacting the data availability policy[13], in force throughout the PLOS series of publications. PLOS instructions to reviewers currently place the burden of checking data availability on the reviewers who may or may not verify the issuance of accession numbers. To mitigate this problem, a watch procedure has been established. It exploits both PLOS One alerts as well as Biocrates' maintained list[14] of scientific publications for which Biocrates AG products appear to notify curators. The demonstration of the overall operational procedure resulted in the deposition of 9 datasets, associated with the following publications:

[10] http://www.cosmos-fp7.eu
[11] http://www.biocrates.com/products/research-products/absoluteidq-p400-hr-kit
[12] https://github.com/ISA-tools/isa-api/blob/master/isatools/convert/biocrates2isatab.py
[13] http://journals.plos.org/plosone/s/data-availability
[14] "Publications - Biocrates Life Sciences AG." http://www.biocrates.com/resources1/publications. Accessed 22 Jun. 2017.

10.1371/journal.pone.0093148, MTBLS251[15]
10.1371/journal.pone.0043764, MTBLS252[16]
10.1371/journal.pone.0074705, MTBLS258[17]

Beyond PLOS publications, the PhenoMeNal initiative led by the University of Oxford has spurred other groups to deposit Biocrates datasets to MetaboLights (Keun et al. MTBLS231MTBLS232 (access restricted); Fiamoncelli et al. MTBLS254[18]; Mapstone et al. MTBLS72[19]; Draisma et al. MTBLS192 (access restricted). Access to these datasets are currently embargoed and will be publically available post publication.

This demonstrates the value of data deposition for reuse, by focusing on interoperation, engagement with vendors and stakeholders. Furthermore, it evidences the specific activities that the PhenoMeNal team are undertaking as part of their important outreach goals as well as contribution to making data FAIR[20].

3.2 Scientific Data Journal

Springer Nature's *Scientific Data*[21] journal focuses on descriptions of data sets provided in machine readable form, relying on the ISA model, complemented by a textual narrative. This constitutes the main article type called a 'Data Descriptor'. A Data Descriptor includes the methods used to collect the data as well as the technical analyses supporting the quality of the measurements. Data Descriptors focus on helping others reuse data, rather than testing hypotheses, or presenting new interpretations, methods or in-depth analyses. Relevant datasets must be deposited in appropriate public repositories prior to the submission of the Data Descriptor. The completeness of these datasets is considered during the editorial evaluation and peer-review processes. Datasets must be made publicly available without restriction in the event that the Data Descriptor is accepted for publication (excepting reasonable controls related to human privacy issues or public safety).

---

[15] http://www.ebi.ac.uk/metabolights/MTBLS251
[16] http://www.ebi.ac.uk/metabolights/MTBLS252
[17] http://www.ebi.ac.uk/metabolights/MTBLS258
[18] http://www.ebi.ac.uk/metabolights/MTBLS254
[19] http://www.ebi.ac.uk/metabolights/MTBLS72
[20] "The FAIR Guiding Principles for scientific data management ... - Nature." 15 Mar. 2016, https://www.nature.com/articles/sdata201618?utm_source=feedburner&utm_medium=feed&utm_campaign=Feed%3A+sdata%2Frss%2Fcurrent+(Scientific+Data). Accessed 22 Jun. 2017.
[21] http://www.nature.com/sdata

In addition to the requirement that datasets are deposited in appropriate data repositories, they must use appropriate data standards. Scientific Data recommends a set of repositories and standards across a range of domain areas (biological sciences, health sciences, chemistry and chemical biology, earth and environmental sciences, physics, astrophysics and astronomy, social sciences). In relation to PhenoMeNal, the subset of biologically-related repositories and standards is of most relevance. A collection of those recommended by Scientific Data can be found at BioSharing[22], which recently broadened its scope and became FairSharing. MetaboLights is the recommended repository for metabolomics data.

Every Data Descriptor published by *Scientific Data* includes a machine-accessible metadata file. This metadata record provides a structured description of the dataset, including key features of the experimental samples and the techniques used to generate the data. Metadata is captured and distributed in ISA-Tab format, which is designed to capture descriptions of research data across disciplines. The Oxford team has developed the ISA-explorer tool to browse, search and visualise the published data descriptors[23]. The tools in PhenoMeNal are being developed to support ISA formats for importing data into the PhenoMeNal infrastructure, and also for exporting analyses. Supporting standardised formats, as recommended by *Scientific Data*, and in particular the ISA format, allows for easy deposition of data that can then be referenced in Data Descriptor manuscripts submitted for consideration and review in *Scientific Data*. For the full submission guidelines, see http://www.nature.com/sdata/publish/submission-guidelines

3.3 GigaScience

The Open University Press/BGI *GigaScience* journal aims to revolutionise reproducibility of scientific analyses by being an open-access, open-data journal, publishing quality assured research objects (data, tools and workflows) across the range of life and biomedical sciences. To this end, the journal publishes standard manuscripts alongside associated data. Like with *Scientific Data*, *GigaScience* articles describing data are asserted a specific manuscript type called ‘Data Notes’ manuscripts, roughly equalling Scientific Datas ‘Data Descriptor’ manuscript type. These Data Note manuscripts can link to quality controlled public repositories, but also the journals own data repository GigaDB, which defines a dataset as a group of files (e.g., sequencing

[22] https://biosharing.org/recommendation/ScientificData.
[23] http://scientificdata.isa-explorer.org

data, analyses, imaging files, software programs) that are related to and support a Data note manuscript or study. Through their association with DataCite, each dataset in GigaDB is assigned a DOI that can be used as a persistent standard citation for future use of these data in other articles by the authors and other researchers. This allows the data itself to be citeable. As with *Scientific Data*, metadata can be captured and distributed in ISA-Tab format for GigaScience submissions.

As part of a collaboration with the Oxford team, the contents of the GigaDB database will be exported in ISA-format and enabling the use of the ISA-explorer tool to browse, search and visualise the data available in GigaDB. For more detail, see http://gigadb.org/site/about

The PhenoMeNal infrastructure can make full use and derive benefit from both Springer Nature's *Scientific Data* and OUP/BGI *GigaScience* journals through their support of the ISA format. PhenoMeNal's use and export of metadata in ISA formats as well as raw and derived data in standard formats (i.e. mzML and nmrML), being developed as outputs of WP8, provide the required interoperability elements to link to the journals.

4 WORK PLAN

In addition to the activities and progress with a series of leading publishers described above, we have begun to contact other publishers based on our extensive analysis of the literature and the compilation of a list of journals that have published in the field of metabolomics[24]. These publishers fall into one of two groups:

- **Group 1** encompasses publishers that already promote the use of metabolomics data repositories and an open access philosophy
- **Group 2** are those that have published metabolomics data but have not yet adopted a policy for making said data open access.

We have written two letters, each to target one of these two groups. The letter to group 1 not only makes the publisher aware of the services provided by the PhenoMeNal e-infrastructure but additionally requests the publisher considers *requiring* the submission of metabolomics data to an open-access data repository as a prerequisite of publishing (see Annex 7.2). For group 2, we request that the publisher considers *recommending* the submission of metabolomics data to an open-access data repository, and our letter also increases their awareness in PhenoMeNal that can standardise the data

[24] For Current endorsements by other journals: https://biosharing.org/biodbcore-000168

generation process by pipelining processing tools into citable re-runable workflows (see Annex 7.3).

We will pursue the following strategy:

- Letters will be sent (August / September 2017) to the editors of the journals compiled in Annex 7.1.
- We will follow up by phone calls where we don't receive responses.
- We will engage with publishers and listen to their requirements.
- By the end of the PhenoMeNal project we expect to have agreements with a further 2-3 publishers to support metabolomics data deposition.

5 DELIVERY AND SCHEDULE

The deliverable is submitted on time

6 CONCLUSION

We have made a focused effort with a select few publishers and initiated a much broader effort with a wide range of journal publishers to raise awareness about and implement our deposition systems. We are trying to model this after *Scientific Data* and *GigaScience* journals.

7 ANNEXES

7.1 List of journals that have published metabolomics

Journals with metabolomics links

Journal	Publisher	Editor	Impact Factor
Metabolomics	Springer	Royston Goodacre: Roy.Goodacre@manchester.ac.uk	3.692
Journal of Proteome Research	ACS Publications	John R. Yates, III: eic@jpr.acs.org	4.268
Molecular and Cellular Proteomics	American Society for Biochemistry and Molecular Biology	Alma Burlingame: alb@cgl.ucsf.edu	6.540
OMICS: A journal of integrative biology	Mary Ann Liebert, Inc., publishers	Vural Özdemir	2.723

Common Journals

Journal	Publisher	Editor	Impact Factor
Nature	Nature Publishing Group (NPG)	Philip Campbell	40.137
FEBS	FEBS Press	Seamus Martin: martin@febs.org	4.237
Current opinion in biotechnology	Elsevier B.V.	Greg Stephanopoulos	9.294
Accounts of Chemical Research	ACS Publications	Cynthia J. Burrows: eic@acr.acs.org	20.268
PLoS ONE	PLOS	Joerg Heber: jheber@plos.org	2.806
Analytica Chimica Acta	Elsevier B.V.		4.950

Clinica Chimica Acta	Elsevier B.V.	Joris Delanghe Alan H. Wu	2.873
Molecular biology reports	Springer	Jonathan Brody	1.828

NMR and Analytical Journals

Journal	Publisher	Editor	Impact Factor
Magnetic Resonance in Chemistry	John Wiley & Sons	Roberto R. Gil Gary E. Martin	1.179
Magnetic Resonance in Medicine	John Wiley & Sons	Matt A. Bernstein	3.924
Analytical Chemistry	ACS Publications	Jonathan V. Sweedler: eic@anchem.acs.org	6.320
Analytical Biochemistry	Elsevier B.V.	Arthur Cooper	2.219
Bioanalysis	Future Science	Sankeetha Nadarajah: s.nadarajah@future-science.com	2.673

Medical Journals

Journal	Publisher	Editor	Impact Factor
Genome Medicine	BioMed Central	Christopher Morrey: editorial@genomemedicine.com	7.07
Molecular Cancer	BioMed Central	Christophe Nicot	6.204
Molecular Oncology	FEBS Press	Julio E. Celis: jec@cancer.dk	5.331
Clinical Cancer Research	AACR Publications	Keith T. Flaherty	9.619
BMC cancer	BioMed Central	Dafne Solera	3.362
Diabetes	American Diabetes Association	Martin G. Myers Jr	8.684
Breast Cancer	BioMed Central	Lewis Chodosh:	6.345

Research		editorial@breast-cancer-research.com	
Pathobiology	Karger	Prof. Dr. Bettina Borisch: Laetitia.Bourquin@unige.ch	1.703
Neoplasia	Elsevier B.V.	A. Rehemtulla	5.006
British Journal of Surgery	John Wiley & Sons	J. J. Earnshaw	5.899
Journal of Hepatology	Elsevier B.V.	R. Jalan	12.486

Pharmacological Journals

Journal	Publisher	Editor	Impact Factor
Pharmacology and Therapeutics	Elsevier B.V.	S.J. Enna	11.127
Biochemical Pharmacology	Elsevier B.V	S.J. Enna	4.581

7.2 Letter to journal editors - group 1 (those that already support some open access metabolomics data policy)

Dear Editor,

The EU-funded PhenoMeNal project develops and deploys an integrated e-infrastructure in the field of metabolomics. It covers a range of workflows from data processing to data mining and is targeted to be suitable for clinical metabolomics, but could be applied to other areas as well, e.g. environmental, biotechnology or plant molecular phenotyping.

Part of our dissemination strategy is to work with publishers to give authors of manuscripts the option to persistently store standardised metabolomics data in readily accessible formats in dedicated data repositories. An example would be the MetaboLights database (http://www.ebi.ac.uk/metabolights/) run by the European Bioinformatics Institute (EMBL-EBI). Persistent storage of metabolomics data will promote greater transparency in, for example, data quality and metabolite identification, and will facilitate the re-use (and associated citations) of the original dataset.

As *your journal* already promotes the use of metabolomics data repositories, we would like to make you aware of the services provided by our e-infrastructure and request you consider *requiring* the submission of metabolomics data to an open-access data repository as a prerequisite of publishing with you.

As we believe in open data as a key measure to foster data re-use, reproducibility and science of high quality, we would like to increase awareness that workflow systems such as PhenoMeNal/Galaxy can standardise the data generation process by pipelining processing tools into citable re-runable workflows.

Irrespective of whether your journal decides upon mandatory data deposition requirements, we would like to build specific links and document procedures for FAIR (Findable, Accessible, Interoperable, Re-usable) metabolomics data deposition. We would be happy to help in this process.

Sincerely Yours,

7.3 Letter to journal editors - group 2 (those that do not currently support any open access metabolomics data policy)

Dear Editor,

The EU-funded PhenoMeNal project develops and deploys an integrated e-infrastructure in the field of metabolomics. It covers a range of workflows from data processing to data mining and is targeted to be suitable for clinical metabolomics, but could be applied to other areas as well, e.g. environmental, biotechnology or plant molecular phenotyping.

Part of our dissemination strategy is to work with publishers to give authors of manuscripts the option to persistently store standardised metabolomics data in readily accessible formats in dedicated data repositories. An example would be the MetaboLights database (http://www.ebi.ac.uk/metabolights/) run by the European Bioinformatics Institute (EMBL-EBI). Persistent storage of metabolomics data will promote greater transparency in, for example, data quality and metabolite identification, and will facilitate the re-use (and associated citations) of the original dataset.

As *your journal* has in the past published metabolomics data, we would like to make you aware of the services provided by our e-infrastructure and request you consider *recommending* the submission of metabolomics data to an open-access data repository as part of publishing with you.

As we believe in open data as a key measure to foster data re-use, reproducibility and science of high quality, we would like to increase awareness that workflow systems such as PhenoMeNal/Galaxy can standardise the data generation process by pipelining processing tools into citable re-runable workflows.

Irrespective of whether your journal decides upon optional data deposition requirements, we would like to build specific links and document procedures for FAIR (Findable, Accessible, Interoperable, Re-usable) metabolomics data deposition. We would be happy to help in this process.

Sincerely Yours,