PhenoMeNal

Deliverable 8.4.1

Project ID	654241
Project Title	A comprehensive and standardised e-infrastructure for analysing medical metabolic phenotype data
Project Acronym	PhenoMeNal
Start Date of the Project	1st September 2015
Duration of the Project	36 Months
Work Package Number	8
Work Package Title	Data provenance, compliance, and integrity
Deliverable Title	D8.4.1 Specifications for derived data matrices specifications and terminology for description of analysis and statistical results
Delivery Date	M24
Work Package leader	UOXF
Contributing Partners	**UOXF, CEA, ICL**, University of Birmingham, EMBL-EBI, IPB
Authors	Philippe Rocca-Serra, David Johnson, Susanna Sansone, Namrata Kale, Reza Salek, Kenneth Haug, Etienne Thévenot, Tim Ebbels,Marta Cascante, Jianliang Gao, Steffen Neumann, Daniel Schober, Payam Emami.

Abstract:

This deliverable provides details on progress towards a regularised reporting of data matrices generated by statistical analysis with an emphasis on univariate multivariate methods and of relevance to users of PhenoMenal computational workflows. The goal is ensure harmonisation and FAIR description of analysis results, allowing quality assurance, unambiguous data reposition and preservation.

Table of Contents

1 EXECUTIVE SUMMARY

The H2020 PhenoMeNal e-infrastructure project aims to deliver a scalable, robust and standards-compliant infrastructure for clinical phenotyping by means of metabolomics techniques.
The purpose of this deliverable, in the context of Task 8.4, is to provide data type specifications for the reporting of metabolomics studies results, i.e. as generated by applying PhenoMeNal Galaxy workflows. These are usually presented in the form of tables of results, also known as data matrices or data cubes. However, those results data matrices come in all forms and shapes, with very little harmonisation on their content. Besides hindering tool-handshake at the result-end of workflow pipelines, this state of affairs induces significant curation efforts for teams operating the public repositories to reliably and consistently archive findings based on metabolomic techniques and researchers aiming to re-analyse and re-use such data matrices.
Furthermore, due to the limiting amount of semantic markup, a range of ambiguity remains as to the nature of measurements being reported and what these are about. Besides quality control, the lack of computer-readable specifications impairs cross-study findability, accessibility and interoperability. Finally, owing to the heterogeneity in structuring information, reuse by processing or integration tools is low. By supporting the FAIR principles[1], data representations with a formal semantics and widely used exchange model will allow for computer-assisted content interpretation/verification and enhance the ultimate value of the generated metabolomics data.
This deliverable, tied to two main sub-deliverables, due on month 24 and month 30 respectively, should be viewed as a progress report on our data standardisation efforts, outlining the current state, future goals and means to achieve them. These goals and detail plans will be described in more depth in upcoming dedicated reports.

This deliverable follows deliverable 8.4, which addressed the reporting in chemical identification and annotation at individual sample levels.

The focus of the present deliverable is to provide a standardised framework for reporting statistical analysis results such as differential analysis using parametric and non-parametric methods (e.g. T-test, ANOVA or Wilcoxon rank-sum test, Kruskal-Wallis test...) for comparing effects of treatments. Besides the regularisation of the data structure, the aim is also to achieve explicit reporting of the statistical results beyond the simple p-value. The reliance on a robust semantic framework (STATO: https://github.com/ISA-tools/stato) for the description of statistics is a central part of the work.

[1] "The FAIR Guiding Principles for scientific data management and ...." https://www.ncbi.nlm.nih.gov/pubmed/26978244. Accessed 24 Apr. 2017.

2 DETAILED REPORT OF THE DELIVERABLE

2.1 Structuring analytical Data Matrices referenced in ISA

2.1.1 Background

Metabolomics data analysis is carried out with a range of tools as shown by the collection of modules available from PhenoMeNal App library. However, each component generate its own output, slightly different from another one, even though they may be handling similar data types and doing similar tasks. Such heterogeneity in data structure and format hampers data flow and an infrastructure such as PhenoMeNal would benefit from regularization in that space.

In order to survey the levels of needs in output standardisation, a compatibility matrix has been produced, listing each tool's data input and output requirements: https://goo.gl/TDEJgx

One needs to distinguish several situations where data matrix harmonisation is needed. EMBL-EBI requires Metabolite Assignment Files (MAF) to be reported. However, annotation requirements differ depending on the analytical technique used.

- Use case 1: How to report Metabolite Identification and Annotation?
- Use case 2: How to report Univariate Analysis?
- Use case 3: How to report analysis results from stable isotope resolved metabolomics?

2.1.2 Implementation

While use case 1 has been covered in earlier work, we report here that an extension to the MAF specification has been made to cover the needs inherent to annotation based on NMR signals. Working with Imperial College London (Tim Ebbels, Jianliang Gao), BATMAN R has been modified to generate an NMR extended MAF file, thereby streamlining the data deposition process for the MetaboLights repository.
To provide a formal description of the data matrices, we chose to rely on the JSON data package[2], a project led by the Open Knowledge Foundation, whose goal is to make more data available and machine accessible. A growing number of projects are defining data packages for their needs and software support exists in key data science languages, such as R and python, with the following components worth noticing, https://cran.r-project.org/web/packages/dpmr/ and https://pypi.python.org/pypi/datapackage respectively
Two specific JSON data packages have also been devised and are available from ISA-tools project GitHub:

[2] A data package is a simple container format for describing a coherent collection of data in a single 'package'. It provides JSON based metadata on tables/schemas and is the basis for convenient delivery, installation and management of datasets. (https://specs.frictionlessdata.io/data-package/). Bindings exist for many programming languages, including Python and R.

https://github.com/ISA-tools/isa-matrix-datapackages/tree/master/src/maf_datapkg

These representations provide a stepping-stone for more complete representations and contain a complete semantic markup, made possible by use of STATO terms. The STATO ontology has been augmented to accommodate specific needs as identified by the process (e.g. addition of 'credible interval', which in a Bayesian context is akin to 'confidence interval').
The new terms will appear in the coming release of the STATO ontology, scheduled for September 2017.

To answer the problems defined by use case 2 and 3, seven different types of JSON data packages have been defined to stabilise and normalise how analysis results should be generated from analysis workflows.
Working with the CEA team (W4M), several metadata profiles have been generated. All are fully semantically marked up, therefore natively making datasets presented in such format compliant with the FAIR principles[3], especially for 'interoperability' and 'reusability'.
The interaction with the CEA team led to a refinement of how to report contrasts (comparison between experimental conditions) and convergence towards the notion of 'effect size estimator' to be used across conditions (instead of specific variants 'diff' for 'means difference' or 'cor' for 'Pearson's correlation coefficient' as used currently).

Using the Frictionless Data Package specification (http://frictionlessdata.io/guides/data-package/), WP8 has generated several metabolomics specific extensions to enable robust definition of data matrices formats, which could then be implemented and established by PhenoMeNal developers in the analysis workflows.

In particular, one can distinguish six main Data Package profiles to address each of the following needs as met under PhenoMeNal conditions:

1. Reporting of metabolite **identification using MS techniques in metabolite profiling**.
2. Reporting of metabolite **identification using NMR techniques in metabolite profiling**.
3. Reporting of metabolite **identification using MS techniques in stable isotope resolved metabolomics.**
4. Reporting of metabolite **identification using NMR techniques in stable isotope resolved metabolomics.**
5. Reporting **univariate analysis results.**
6. Reporting **multivariate analysis results**.

[3] "The FAIR Guiding Principles for scientific data management and ...." 15 Mar. 2016, http://www.nature.com/articles/sdata201618. Accessed 29 Aug. 2017.

```
{
    "name": "q-value",
    "title": "q-value",
    "description": "adjusted p-value using a false discovery rate correction method",
    "format": " default",
    "type": "number",
    "rdfType": "http://purl.obolibrary.org/obo/OBI_0001442",
    "constraints": {"required": "False"}
},
{
    "name": "mlt_corr_method",
    "title": "multiple testing correction method",
    "description": "multiple testing correction method to be selected from a controlled list of value",
    "format": " default",
    "type": "string",
    "rdfType": "http://purl.obolibrary.org/obo/OBI_0200089",
    "constraints": {
        "required": "False",
        "enum": [
            "Bonferroni",
            "Holm-Bonferroni",
            "Benjamini and Hochberg",
            "Benjamini and Yekutieli"
        ]
    }
},
{
    "name": "95pc_upr_bound",
    "title": "upper confidence bound of a 95% confidence interval",
    "description": "upper confidence bound of a 95% confidence interval qualifying the statistics",
    "format": " default",
    "type": "number",
    "rdfType": "http://purl.obolibrary.org/obo/STATO_0000314",
    "constraints": {"required": "False"}
},
{
    "name": "95pc_lwr_bound",
    "title": "lower confidence bound of a 95% confidence interval",
    "description": "lower confidence bound of a 95% confidence interval qualifying the statistics",
    "format": " default",
    "type": "number",
    "rdfType": "http://purl.obolibrary.org/obo/STATO_0000315",
    "constraints": {"required": "False"}
},
{
    "name": "means_diff",
    "title": "difference between means",
    "description": "the difference between the means computed over each of the group involved in the comparison/contrast",
    "format": " default",
    "type": "number",
    "rdfType": "http://purl.obolibrary.org/obo/STATO_0000085",
    "constraints": {"required": "False"}
},
```


Figure1: An excerpt from the *phnml-t-test-group-comparison-datapackage.json* showing how data matrix field headers are formally defined. The JSON description requires the specification of a field name, a description which is used in rendering components to provide contextual help, a type which specifies the data type expected for that field, thus enabling validation. Finally, the semantic markup is provided via the *rdfType* element. The associated value is a persistent, resolvable URI to classes in the relevant semantic frameworks, in this excerpt, in classes from STATO and OBI.

In total, 9 JSON data package extensions and metadata profiles have been produced and are currently being tested (see Figure 1 for an example). They are available from a dedicated Github repository under the ISA-tools project:
https://github.com/ISA-tools/isa-matrix-datapackages

These data packages cover cases 1,2 and 3. This last case covers needs expressed by the community of users whose efforts focus on stable isotope resolved metabolomics. Further interaction is needed to work towards implementation but it is worth noting that the layout and structure offered by the corresponding JSON data packages for SIRM data aligns closely with

the native output from tools such as Ramid or iso2flux. It is therefore anticipated the uptake will be quick.
The case of multivariate analysis necessitates additional work and will be the focus of further efforts both on the structuring and extension of the semantic framework needed to support this use case.

In parallel, additional work has been carried out on a tabular format for representation of quantification and identification of metabolites by repurposing and extending mzTab format currently being developed by the HUPO-PSI initiative and members of the metabolomics community. A meeting held at the EBI at the end of August 2017 brought representatives of IPB (Steffen Neumann), EMBL-EBI (Kenneth Haug, Reza Salek, Claire o'Donovan), HUPO-PSI (Andrew Jones) in order to review the draft specifications of the mzTab format for Metabolomics (https://github.com/HUPO-PSI/mzTab/tree/master/specification_document/1_1_draft_specs). Several discussions about the overlap with the existing MAF files and JSON data packages have been raised as well as the issue of implementation. The work will be pursued as to resolve the ambiguities and ensure implementation in the H2020 PhenoMeNal workflows.

3 WORK PLAN

Consistent with the work so far and building on it, WP8's attention for deliverable D8.4 has been focused on collecting the needs from the community and practitioners, regularly meeting with them and reaching out for input in order to shape specifications attuned to real use cases. The actual delivery of the standardisation format and supporting documentation is to take place in the remaining tasks and deliverables (T8.4 and D8.4.1-D8.4.2, delivered in months 24 and 30 respectively).

Objectives

O8.1 Define metadata and data exchange standards, along with technical and user documentations.

O8.2 Implement and maintain PhenoMeNal reference implementations.

Tasks

T8.1: Use cases and state of the art of communication standards

T8.2: Standards for exchanging experimental and clinical metadata

T8.3: Data standards exchange formats

T8.4: Harmonisation of data matrices and analytical results

T8.5: Maintain documentation and disseminate information

Deliverables

D8.1 Report on community standards for reporting, access and integrity supported in the PhenoMeNal grid; to be disseminated in a dedicated BioSharing page and via the project website. (M12)

D8.2: Modularized ISA model and format: biospecimen centric schema, corresponding xml schemas, reference implementation guidelines and validation rules. (M24)

D8.3: nmrML, mzML data exchange formats and associated terminologies for instrument raw, with reference implementation guidelines and validation rules. (M18)

D8.4: Signal processing and analysis data exchange format

D8.4.1: Specifications for derived data matrices, specifications and terminology for description of analysis and statistical results (M24)

D8.4.2: Reference implementation guidelines and validation rules (M30)

The next key deliverables therefore are:

D8.4.2: Reference implementation guidelines and validation rules (M30)

4 DELIVERY AND SCHEDULE

The delivery is delayed: No

5 CONCLUSION

By recommending and using the data package specification to describe result data matrices in a human and machine readable format, we not only clarify the conditions for creating consistent data matrices, but also provide a method that allows a complete semantic markup of the information held in data matrices generated by H2020 PhenoMeNal tools, therefore ensuring complete interoperability and reusability of the data. For example, the dedicated data package specifications developed for this deliverable now allow extensive provenance tracking and hopefully will ease data pipelining across tools and workflow modules and information reporting compliance.

Furthermore, the presented work has the potential to result in a significant improvement in the quality of dataset reporting in the context of deposition to public repositories, i.e. when using PhenoMeNal infrastructure for carrying out classic statistical exploration of metabolomics datasets.

Finally, the extensive semantic markup of the data matrices means that it enables seamless conversion of tab delimited data tuples to RDF triples, therefore paving the way to expose metabolomics data as semantic web representations under the Linked Open Data (LOD, http://linkeddata.org/) cloud and meeting the highest threshold of requirements for ***FAIRification***.