
## Deliverable 8.3

| | |
|---|---|
| **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** | D8.3 nmrML, mzML data exchange formats and associated terminologies for instrument raw data, with reference implementation guidelines and validation rules |
| **Delivery Date** | M18 |
| **Work Package leader** | UOXF |
| **Contributing Partners** | IPB, UOXF, CIRMMP, UB |
| **Authors** | Daniel Schober, Steffen Neumann, Philippe Rocca-Serra, David Johnson, Antonio Rosato, Marta Cascante, Reza Salek |

**Abstract:**
This deliverable reports on the creation of a new open NMR data standard and the update of another one for Mass Spectrometry to support the exchange and deposition of raw instrument NMR and MS data used for Metabolomics applications.
Furthermore, the report details how these formal specifications are now supported by software tools and implementations, facilitating format uptake and use by academics and manufacturers alike, linking to documentation and user guides.

## TABLE OF CONTENTS


1. EXECUTIVE SUMMARY ..... 3
2. CONTRIBUTION TOWARDS PROJECT OBJECTIVES ..... 4
3. DETAILED REPORT OF THE DELIVERABLE ..... 6
   3.1. Data standards for metabolomics research ..... 6
   Data Standards Survey ..... 7
   Assess data standards and terminologies for the compute pipelines ..... 7
   3.2. Standards for NMR metabolomics data ..... 9
   The nmrML open access data standard for NMR raw data ..... 9
   3.3. Standards for mass spectrometry metabolomics data ..... 13
   mzML as open mass spectrometry raw data standard ..... 14
4. WORK PLAN ..... 15
5. DELIVERY AND SCHEDULE ..... 16
6. CONCLUSION ..... 17
   7. ANNEX ..... 18
   Annex 1: FAIR criteria ..... 18
   Annex 2: Software producing and consuming nmrML ..... 18
   Annex 3: nmrML semantic validator example ..... 21
   Annex 4: nmrML Git statistics ..... 22

# 1. EXECUTIVE SUMMARY

The H2020 PhenoMeNal e-infrastructure program aims to deliver a scalable, robust and standard-compliant infrastructure for clinical phenotyping by means of metabolomics techniques. One of the PhenoMeNal objectives is to establish the technology for an audit trail for the processing of human metabolic phenotyping data from the raw data acquisition all the way to the generation of high-level biomedical insights. An important contribution towards these goals is to propose traceable data representation formats that serve to standardize the communication channel between the data processing modules within the pipeline. The aim of this deliverable is to provide an informed selection of suitable open access data standards serving a wide range of workflow modules for NMR and Mass Spectrometry data processing. This selection was guided by a) the requirements of the processing modules in our use case workflows (data driven standards selection), b) by the outcome of the previous 'D8.1 Report on community standards', and c) by looking at the answers provided in an additional metabolomics workflow survey (http://link.springer.com/article/10.1007/s11306-016-1147-x) launched by the University of Birmingham via ELIXIR-UK and the Metabolomics Society, with participation from PhenoMeNal partners. The outcome of our analysis was that the most prominent open access data standards are mzML for the mass spectroscopy domain and nmrML for the nuclear magnetic resonance domain. This selection was also driven by the fact that mzML and nmrML are the recommended open standards as supported by the domains standardisation governance bodies (PSI http://www.psidev.info/ and MSI http://www.metabolomics-msi.org/ respectively). We have to mention that the nrmML standard is already a release candidate, and we contributed to this open NMR raw data standard, which was build around well established technologies. To this end, an approach similar to that used by the HUPO-PSI community to develop the mzML and PSI-MS controlled vocabulary has been adopted. Build around the widely accepted XML technology, an XML schema definition (XSD) for NMR data has been created and is now published as nmrML (NMR markup language, www.nmrML.org). Along with the nmrML.XSD, we release a supporting controlled vocabulary (CV) called nmrCV, which can be referenced from within nmrML files for further annotation with standardized terminological descriptors. In order to foster ease-of-adoption, we have developed format converters that generate valid nmrML raw data files from vendor specific raw data formats. These converters have been containerized for inclusion into PhenoMeNal workflows. We have also developed validation services that allow to validate user- or autogenerated nmrML files and CV terms used therein with regard to coverage completeness. The validator containers also detect errant element nesting, hence contributing to quality assurance on the data

representation side. nmrML and its accompanying CV were started in the COSMOS EU project (http://cosmos-fp7.eu) and is now being continually updated within PhenoMeNal by an interdisciplinary team of experts from around the world, including academics but also involving representatives from the instrument manufacturer Bruker Biospin.
For mass spectrometry data, we designed tools and pipelines to accept the mzML format, as this format is already mature and well established in the community. We also contribute to the mzML development and maintenance by proposing missing elements or terms required for novel instruments and applications used in metabolomics.

## 2. CONTRIBUTION TOWARDS PROJECT OBJECTIVES

Thousands of articles using metabolomics approaches are published every year. With the increasing amounts of data being produced, mere description of investigations as text in manuscripts is not sufficient to enable re-use anymore. Ideally, the data needs to be published together with the findings in the literature, along with the process detailing how the metabolomics data is processed. These should be published in a standardized form[1] to aid quality assurance, enable secondary data usage and ultimate contribute to avoid a waste of public and private expenditure.

One of the PhenoMeNal objectives is to establish the technology for a robust audit trail for the processing of human metabolic phenotyping data from the raw data acquisition all the way to the generation of high-level biomedical insights. The outcome of this deliverable assures that open access data standards can be readily used as information channel between the required data processing containers within metabolomics workflows. Standards are developed to ensure that scientific information is delivered consistently, efficiently and meaningfully to the benefit of the community and processing algorithms alike, ultimately contributing to intelligible audit trails within workflows. By leveraging on open access formats, we aim to increase accessibility, reproducibility and interpretability of the data that is passed through PhenoMeNal workflows.
To re-run workflows, i.e. to allow secondary data usage or further validation, it is vital to precisely track every step of the processing workflow of a study, allowing anyone to get reliable insight at any level of the study. Currently, scientists most often struggle with repeating these steps due to the vast amount of heterogenous descriptors and data formats applied. That is why in PhenoMeNal we encourage the use of data standards: In recent years, the notion of FAIR (**F**indable, **A**ccessible, **I**nteroperable and **R**eusable) research data objects has been endorsed by an increasing number of researchers and organisations, including the Dutch Techcenter for Life Sciences (DTL,

[1] "Data standards can boost metabolomics research, and ... - SpringerLink." http://link.springer.com/article/10.1007/s11306-015-0879-3. Accessed 12 Feb. 2017.

http://www.dtls.nl/) and the FORCE11 (https://www.force11.org) and Data FAIRport (http://datafairport.org/) initiatives. Data standards help to make data FAIR (see Annex 2) and contribute to the Open Access philosophy. Increasingly, funding agencies are requesting publicly-funded research data to become *open access*. Recent calls for making publicly funded data be publicly available has resonated loudly and many groups including PhenoMeNal are weighing-in to end data retention by scientists[2,3].
The terminological standardisation, which is established by using ontologies in amendment to the XML based data standards, will contribute to alignment of verbal expressivity, which will in turn foster coherent semantic interpretation of the data annotated with these ontological expressions. This in turn can considerably ease the positive recall in data queries and can increase precision and lower the amount of false positives in search-result lists.
By introducing data validation schemes, we can assure that data audit trails become sufficiently complete and robust. By regularising how information is structured and reported, standards make it easier to reproduce, distribute, and exchange information as it narrows down the expressive variances. Together, these quality aspects will ultimately aid to make the generated data FAIR (http://www.nature.com/articles/sdata201618) for the workflows and for long term storage in repositories, hence also increasing its chances for being exploited for secondary data analysis by 3rd parties.

As part of work towards Deliverable D8.4, whose goal is to propose formalized data matrices resulting from analytical processing, UB and UOXF have been gathering use cases, reviewing literature and existing practices of data reporting in the specific application of the metabolomics techniques in the context of stable isotope resolved metabolomics. The work has already lead to the implementation of the directives into tools for mass spectrometry based applications (phnmnl/container-midcor, phnmnl/container-ramid, and phnmnl/container-isodyn). For NMR derived data, further interactions with domain experts in leading teams in Europe and the USA is taking place (NIH centers meeting Sun Feb 26, 2017, San-Diego). The outcome of the work will be the basis for evolving nmrML standards where needed to ensure the standard can accommodate the needs, and thus contribute to the FAIRification of stable isotope resolved metabolomics studies

Our activities have hence contributed to the following Objectives as specified in the DOW:
**8.1** Define metadata and data exchange standards, along with technical and user documentations.

[2] Free the Data Activity by Genetic Alliance <http://www.free-the-data.org/>
[3] J C Molloy “The Open Knowledge Foundation: Open Data Means Better Science” PLoS Biology, 9(12): e1001195

**8.2** Implement and maintain PhenoMeNal reference implementations.

# 3. DETAILED REPORT OF THE DELIVERABLE

## 3.1. Data standards for metabolomics research

In recent years, metabolomics data standards have developed extensively, to include the primary research data, derived results and importantly the experimental metadata in a machine-readable way. In our case this includes vendor-independent open access data standards such as mzML for mass spectrometry (http://psidev.info/mzml_1_0_0%20) and nmrML for NMR raw data (http://nmrml.org) that have fostered the development of advanced data processing algorithms by the scientific community. The PhenoMeNal approach will contribute to such formalization and reproducibility by fostering the use of accepted open access standards for the representation of metabolomics data (by mzML and nmrML), experimental metadata (by ISA-Tab) and workflow data (via Galaxy). The efforts to standardize the experimental metadata will be described in our upcoming 'D8.2 Modularised ISA model', and the standards to propose for the end of our metabolomics pipelines will be described in the future 'D8.4 Signal processing and analysis data exchange format'.

When proprietary software is being used which does not support open standards, narratives are needed to document the representations involved, i.e. parameters used and user interactions with the graphical user interface of the software. Such unstandardized narratives are often not sufficient to reproduce a whole workflow, as informal 'narratives' fail to provide computer-accessible semantics and versioning of the software, graphical user interfaces and default parameters could change over time. In metabolomics, the multitude of different technological platforms (i.e. >30 different mass spectrometry vendors) is an important source of heterogeneity and bias, which can be overcome by leveraging on agreed-upon subsets covering the relevant bits of information in one reduced and aligned, but still sufficiently complete, and community agreed data standard. At the beginning of a workflow, metabolomics studies often comprise of many raw data files, so the conversion from such Instrument-generated vendor formats must not involve expensive manual intervention to add information beyond what is already stored in the instrument software. PhenoMeNal therefore assures that dockerized vendor-to-open-data-standard converters are available for data processing workflows.

Further incentives to prepare standards-compliant data sets include new opportunities to publish data sets, but also require a little "arm twisting" in the author guidelines of

scientific journals to submit the data sets to public repositories such as the NIH Metabolomics Workbench or the MetaboLights at the EBI. Here, we encourage (re-)use of MetaboLights data sets by making a container for Metabolights download available for workflow integration.

Altogether, the usage of data standards in PhenoMeNal workflows will pave the way for both reproducible research and data reuse, including quality and meta-analyses.

**Data Standards Survey**

**Assess data standards and terminologies for the compute pipelines**

The links between tools in a processing pipeline are representations of input- and output-data, which should be standardized where possible; mainly in order to avoid reinventing the wheel, limit parser writing and ensuring community agreed-upon meanings.

Liaising between WP8 and WP9 activities is therefore essential, building on the Initial WP 8 analysis of community standards D8.1[4] and the published BioSharing collection found at https://biosharing.org/bsg-c000028.

In general, we have to distinguish standardization of the data syntax (e.g. all agreeing on a XML or CSV/data matrix tabular data representation); and standardization of the data semantics (e.g. how a syntax-inherent grammar is to be parsed in order to allow the meaning behind the data to be captured in a human intelligible and computer-interpretable form). On the syntax side, most tools accept data matrices as output formats. Tools at the beginning of the workflows also accept XML based open community standards for vendor independent raw data capture (i.e. mzML for Mass spectrometry and nmrML for NMR raw data). On the semantics-side, often ontology-based terms are applied in these standards, which provide the community-agreed meaning of these descriptors used for data annotation. We have checked the involved metabolomics assaying domains (Mass spectrometry, NMR, Fluxomics) and identified the following data standards supported by the open source metabolomics tools:

For **Mass spectrometry,** the mzML XML format[5] and accompanying Controlled Vocabulary (CV) allows to capture instrument raw data in a vendor agnostic fashion. For **NMR,** the nmrML XML format[6] and accompanying CV does the same. An **assay type agnostic metadata standard** is currently being established for higher-level experimental metadata, also allowing for CV-based enrichment: **ISA-Tab**[7]. This feature

[4] http://phenomenal-h2020.eu/home/wp-content/uploads/2016/09/Deliverable8.1.pdf
[5] http://www.psidev.info/mzml_1_0_0
[6] http://nmrml.org/
[7] http://isa-tools.org/

offers the ability to support multi-omics, multi-technology phenotyping studies through its extensibility.

**At the beginning of the workflows**, vendor neutral raw data standards and converters into those open-access formats play a more prominent role, and are part of the preprocessing tools section. In the case of additional experimental metadata capture as intended by the ISA approach, the ISA tools provide a mechanism to look up and add CV terms from ontologies available in the bioportal library. Here, ISA configurations help to ensure the correct ontology is selected and appropriate coverage is provided. Although ontology use is not enforced, recommendations for the use of medical phenotyping ontologies can be issued, i.e. to pick terms from the terminological standards shown in Table 1.

**In the middle of the workflows**, where we find a plethora of variant data processing and enrichment tools, coverage by accepted standards is naturally sparse, and hence usually underspecified data-matrices prevail. Although sometimes published CVs and ontologies cover some of the required descriptors for such middle layer processing tools, the curator-burden on the annotation side has hindered the wider ontology-based annotation. Here, piping CV terms captured at the beginning of the workflows, i.e. in mzML and nmrML raw formats, through into ISA and from here to the processing tools is a way to increase availability of CV terms in workflows.

**At the end of the workflows,** we expect more data standards to be implemented in the future, i.e. to capture basic end-results - the scientific outcomes of a study - which pose the searchable outcome of the overall data reduction and publication process. We here name mzTab [http://www.psidev.info/mztab] for Mass spectrometry. We also investigate quality assurance terminologies such as the Metabolite Identification Evidence Code Ontology (MIECO)[8].

In general, a selection of life-science-related standards is provided in the **BioSharing** library [www.BioSharing.org]. This repository also provides **minimum information (MI) checklists** for most application domains, which can assist in information content coverage assurance. Such MI standards are currently enforceable for nmrML, mzML and ISA data, and for the latter a specific ISA configuration [http://isa-tools.org/format/configurations/] validation ensures mandatory minimal metadata availability as specified in the MSI-sanctioned **Core Information for Metabolomics Reporting (CIMR)**[9].**Table 1:** Terminological Standards delivering terms and descriptors for PhenoMeNal workflow annotations

[8] http://ceur-ws.org/Vol-1692/paperE.pdf
[9] http://mibbi.sourceforge.net/projects/CIMR.shtml https://biosharing.org/bsg-s000175

Currently, as part as of the continued standards watch and interaction with standardization efforts by chief Standard Development Organizations (SDO) (GA4GH, HL7, CDISC…), semantics standards to describe clinical consent and availability, data access and data use conditions, PhenoMenal is evaluating a recently released vocabulary, 'Data Use Ontology' (DUO). In collaboration with WP5 (ELSI), ISA study metadata profiles are being evaluated to include key information. DUO has only recently being publicly released and is being tested by GA4GH, EGA and Broad Institute.

Another area where WP8 is watching is an effort around the regularized description of the workflow itself. The Common Workflow Language initiative aims at delivering a specification that would allow descriptions of computational workflows to become portable and shareable within and accross a wider range of VRE and workflow platforms. This effort is currently engaged in a round of implementation, evaluation and testing with major workflow engines, including Knime, Galaxy, Taverna, Arvados and others.

### 3.2. Standards for NMR metabolomics data

NMR has been extensively used in clinical metabolomics to study metabolites in patients' biofluids and tissue extracts. With a growing number of NMR repositories for metabolomics, demands for a vendor-agnostic, open data format have emerged. This is because data in proprietary formats age quickly and NMR data stored in such formats can often become obsolete, making valuable results inaccessible and irreproducible in the long run[10]. An open and persistent, vendor-neutral data standard is needed for long-term archive and storage of NMR experimental data and this open standard is currently emerging as nmrML[11].

**The nmrML open access data standard for NMR raw data**

We believe the nmrML data standard will ease and encourage sharing, comparison and reuse of NMR data from clinical metabolomics experiments in public data repositories and will become a key element of PhenoMeNal Virtual Research Infrastructure. As instrument vendors typically provide the data processing software and proprietary data formats together with the instrument hardware, developers of third party NMR analysis software often need to devote considerable effort into reading and writing these vendor-specific formats. This applies both to commercial software and to community developed open-source tools such as the Batman R package (1), Bayesil (2), NMRProcFlow (3), rNMR (4) and MetaboAnalyst (5). With the recent termination of the Agilent/Varian NMR spectrometer range and related software support, the question of long-term readability of discontinued vendor formats has become pertinent for a growing NMR community.

[10] http://pubs.rsc.org/en/content/articlepdf/2016/np/c6np00022c
[11] www.nmrML.org

NMR spectra processing and quantification tools will benefit from the standardized nmrML storage.

Figure 1 summarizes available nmrML-compliant tools and functionalities in support of a typical NMR data handling workflow for a metabolomics or similar experiment.

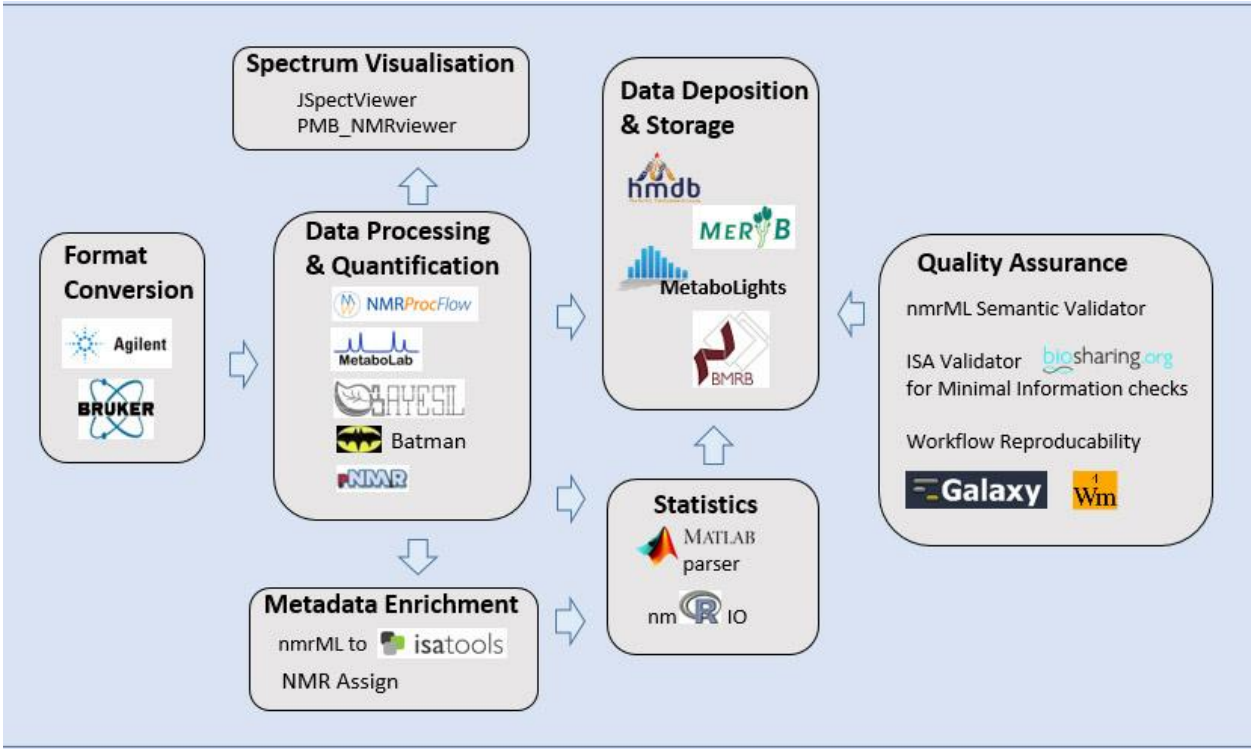



**Figure 1**: An abstracted metabolomics workflow of NMR data processing and storage is shown and nmrML-aware tools supporting each workflow steps are illustrated. Vendor to nmrML converters, NMR data processing and visualization tools, as well as public repositories accept nmrML as standard data format. Parsers for MATLAB(R) and R make nmrML data accessible to statistics tools and content validators assist in data quality control and workflow reproducibility.

We are currently testing the extent to which our NMR-based PhenoMeNal use cases can be supported by nmrML throughout the processing workflow. nmrML is mainly used for 1D raw and some basic processed / annotated NMR data storage throughout the pipeline. We assure that a range of PhenoMeNal workflow modules can take nmrML as raw/FID data input and, where appropriate, produce nmrML outputs augmented with spectral processing data such as compound identification and quantitations. An example of how nmrML captures annotated and quantified molecule lists can be found in Example 4 of the documentation webpage[12]. While the nmrML.xsd covers raw data, it also provides descriptor elements encompassing open source NMR processing and quantification outputs, as generated by Bayesil, Batman or nmrProcFlow[13]. This allows easy integration of modules in computational pipeline consuming nmrML documents. The fact that selected basic metadata is captured within the same file as the raw data

[12] http://nmrml.org/examples/4/
[13] See Annex 2 for a list of nmrML compliant tools

eases pipeline development by reducing the complexity of file tracking within the workflow environment (e.g. Galaxy), as data moves between successive pipeline modules. Further, the modules can be used as standalone tools or re-combined if needed, as the nmrML converters and readers are available as standalone modules. The outputs of the pipeline, or any module, can easily be downloaded and shared between research groups or through public repositories[14]. In addition, the nmrCV allows for a detailed standardized description of NMR workflow functionalities. The overall bulk of study design metadata (Investigation, Study, and Assay context information) however is captured using ISA model and specific metadata profiles defined in ISA configurations.

The nmrML format specification is composed of an XML schema and an companion controlled vocabulary, nmrCV. As a data exchange format, nmrML captures raw NMR data, spectral data acquisition parameters and, where available, spectral metadata such as identified chemical structures associated with spectral assignments and chemical concentrations. Although currently only capturing 1D NMR data, the model already possesses the required additional axis for 2D data capture and plans are laid out to make this feature usable in the next nmrML revision published, after gathering an initial community feedback for our first-release paper publication.

To facilitate automatic format conversions, we have created several nmrML converters (one in Java and an alternative in Python) for Bruker and Agilent/Varian vendor formats. In addition, easy-to-use, web-based spectral viewing, processing and spectral assignment tools that read and write nmrML have been developed. An nmrML semantic validator allows to check for the correct implementation of manually populated nmrML files, i.e. with respect XML schema compliance, CV term usage and allowed term cardinalities. At the core, the XML syntax and structural validity of nmrML XML instances (XML element and attribute positions, order and cardinality) can be checked by any validating XML parser against the nmrML.xsd, which defines these allowed elements and their expected characteristics. On the next layer, so-called mapping rules can enforce semantic validity of the ontological descriptions used, by testing which CV terms are allowed in which elements. The elements with their allowed CV descriptors are outlined in the mapping rule file. The OpenMS/Topp-based nmrML validator (http://nmrml.org/validator/) checks that these higher level semantic criteria are being met in a given XML instance, e.g. a validation rule file can enforce minimal reporting guidelines, i.e. ensure mandatory minimal metadata availability as specified in the MSI-sanctioned Core Information for Metabolomics Reporting (CIMR)[15]. These validation scenarios make nmrML more easily accessible to quality assurance than JCAMP-DX or other more verbose and equivocal formats. An example validation has been added as Annex 3.

[14] MetaboLights accepts nmrML as NMR raw data format
[15] http://mibbi.sourceforge.net/projects/CIMR.shtml

Although this is the first iteration of nmrML for capturing and disseminating 1D NMR data for small molecules, the nmrML format has already been adopted by several open source data processing tools and metabolomics reference spectral libraries, e.g. serving as storage format for the MetaboLights data repository. The nmrML open access data standard has been approved by the Metabolomics Standards Initiative (MSI)[16]. The nmrML core specification, including the XSD and nmrCV, can be found at nmrml.org, together with tutorials. The referenced nmrCV.owl currently contains over 600 terms and is indexed under the NCBO Bioportal ontology library. On our documentation website, we provide code examples for single compound reference spectra as well as mixed-compound NMR spectra.

Development was coordinated via mailing lists, video conferences, and during multiple workshops and hackathons. The choice of XML over JSON, or other recent data formats, was motivated by technical maturity, flexibility and universality of XML in both capturing and presenting scientific data. There is abundant XML expertise and tooling to leverage on, as XML resides at the base of the semantic web stack. We implemented format converter tools to generate valid nmrML from vendor raw data files. Links to nmrML compliant databases as well as NMR processing and spectrum visualisation software are provided. Format parsers, application program interfaces (APIs), and validation webservices have been set up. All code libraries, an issue tracker as well as a file versioning and release policy are available on the nmrML developers GitHub pages at https://github.com/nmrML/nmrML. We have added an overview of the recent nmrML git development metrics as Annex 4.

### 3.3. Standards for mass spectrometry metabolomics data

Early mass spectra were intended for human inspection, initially as images on photo plates, or printed as spectra or peak lists on paper. In the 1990's, the IUPAC CPEP Subcommittee on Electronic Data Standards developed the JCAMP formats[17,18] to harmonise the peak lists and associated spectral metadata in a human and computer readable manner. The human readability had disadvantages as the storage space for the textual representation required a whole byte for each digit. The Network Common Data Form (netCDF) was developed about 25 years ago[19] for data in vector and array representations, such as geospatial data in climate models. The benefits of netCDF, which was optimised for efficient storage and access, lead to the specification of

[16] http://www.metabolomics-msi.org/
[17] http://www.jcamp-dx.org/
[18] Lampen, Peter et al. "JCAMP-DX for mass spectrometry." *Applied spectroscopy* 48.12 (1994): 1545-1552.
[19] Rew, Russ, and Glenn Davis. "NetCDF: an interface for scientific data access." *Computer Graphics and Applications, IEEE* 10.4 (1990): 76-82.

Analytical Data Interchange Protocol for Chromatographic Data[20],[21] (ANDI-MS) which was adopted by the American Society for Testing and Materials (ASTM)[22].

**mzML as open mass spectrometry raw data standard**

About 10 years ago, two separate XML standards were developed independently, mzXML[23] under the guidance of the "mzXML-associated standard solutions" (MASS) Committee, and mzData[24] within the proteomics standardisation initiative (PSI). By 2009, the best aspects of both mzXML and mzData were consolidated and merged into a new standard called mzML[25] and joint support for a single open standard, thus eliminating duplicated efforts.

For all three XML based formats, and already described for nmrML above, the following factors were vital for broad adoption: 1) the support by vendors of MS instruments and the existence of freely available converters from vendor formats to the corresponding XML, 2) the availability of Open Source parser libraries, including validators to ensure completeness, consistency and unambiguous encoding of information. These in turn facilitate: 1) the broad support in Open Source research software and consequently 2) the adoption of mzML by major data repositories such as MetaboLights[26] and PRIDE[27], that both encourage or even enforce data deposition in vendor independent (non-proprietary) formats.

During the PhenoMeNal project we have so far not encountered a case where the existing mzML format was insufficient for the data processing. This applies to all files that were uploaded to the MetaboLights repository. As of February 2017, MetaboLights contained 36.146 MS raw data files in open format, for a total of 3330 GB of data. Additional (and larger amounts) of data are stored at the Phenome centers, which have the same characteristics.

We were in contact with David Heywood from the MS vendor Waters about the requirement to give access to the recalibrated MS raw data for tools using the Waters

[20] http://www.astm.org/DATABASE.CART/HISTORICAL/E1947-98.htm
[21] Erickson, Britt. "Government and Society: ANDI MS standard finalized." *Analytical Chemistry* 72.3 (2000): 103 A-103 A.
[22] "ASTM International - Standards Worldwide." 27 Mar. 2015 <http://www.astm.org/>
[23] Pedrioli, Patrick GA et al. "A common open representation of mass spectrometry data and its application to proteomics research." *Nature biotechnology* 22.11 (2004): 1459-1466.
[24] Orchard, Sandra et al. "Further advances in the development of a data interchange standard for proteomics data." *Proteomics* 3.10 (2003): 2065-2066.
[25] Martens, Lennart et al. "mzML—a community standard for mass spectrometry data." *Molecular & Cellular Proteomics* 10.1 (2011): R110. 000133.
[26] Haug, Kenneth et al. "MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data." *Nucleic acids research* (2012): gks1004.
[27] Vizcaíno, Juan Antonio et al. "The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013." *Nucleic acids research* 41.D1 (2013): D1063-D1069.

raw data access DLLs, and currently Waters is changing the data access in their conversion DLLs to allow that access. This had been an issue that was troubling mzML over many years.

## 4. WORK PLAN

The focus of WP8 up to this deliverable D8.3 has been to engage with the community around metabolomics data and understand the needs of the community to help inform the standards selection and development that is to take place in the remaining tasks and deliverables (T8.1-8.5 and D8.2-D8.4).

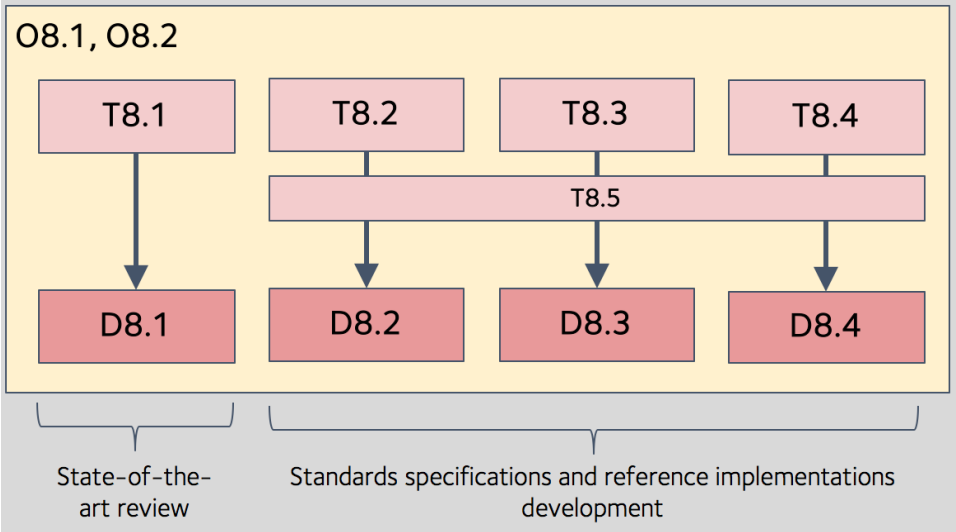



**Figure 2**. Structure of the WP8 workplan of tasks, deliverables and objectives.

### 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**: Harmonization 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:**
**D8.2 Modularized ISA model and format:** biospecimen centric schema, corresponding xml schemas, reference implementation guidelines and validation rules.
**D8.4.1 Specifications for derived data matrices specifications and terminology for description of analysis and statistical results**, both at month 24. Work towards these has already begun (progress can be found in Pivotal tracker), with much work towards the reference implementations and tooling underway in collaboration with WP9.

## 5. DELIVERY AND SCHEDULE

The delivery is delayed: No

## 6. CONCLUSION

The recommendation to use open standards in PhenoMeNal workflows should be perceived and understood as a vehicle to increase trust, secondary usage and higher visibility of scientific output. Reused data is useful data and is more likely to get cited. Standards compliance is just another standard operating procedure applied to the dissemination of the research output in the reproducible science context.

The work accomplished under D8.3 has delivered a key component for establishing a re-useable data processing infrastructure, establishing unambiguous and workflow-friendly data formats for expressing raw and processed data to be consumed and generated by MS and NMR workflow containers. Specifically, we have contributed to the generation of nmrML, an XML-based open standard for 1D NMR data that is leveraging on nmrCV, an aligned controlled vocabulary describing the NMR terminology in a taxonomic way. Such vendor neutral community agreed NMR standard was not available up until now, in spite of several specialized formats focusing on structural information and suffering from verboseness due to dialect establishments (e.g. JCAMP). Work in D8.3 coordinated not only the production of the format specification and companion vocabulary, it also coordinated the delivery of an entire set of software tools supporting the format (see Annex 2). This support ranges from syntax validation to semantic validation (by means of rules) as well as conversion from vendor specific file formats to the PhenoMenal recommended (and MSI sanctioned) open nmrML standard. Leveraging on open standards is expected to increase reusability and interoperability, e.g. along the FAIR principles (Annex 1). nmrML provides a major building block for ensuring regularized input data to the library of operational workflows being developed under PhenoMenal to carry out analytical tasks. Undoubtedly, clean and precise data acquisition and storage is vital and, for that, the utilization of computers for processing, interpretation, statistical analysis, evaluation and reporting of the results has become key[28]. The deterministic nature of computer algorithms means that the same tasks applied to the same data should produce the same output.

Finally, dedicated software components have been developed to automate the creation and synchronisation of ISA documents from/with sets of nmrML or mzML files. Storing the experimental raw and result data along with its ISA metadata is also a prerequisite for long term storage and deposition to dedicated metabolomics repositories like Metabolights.

[28] http://link.springer.com/article/10.1007/s11306-007-0081-3

# 7. ANNEX

## Annex 1: FAIR criteria

| Criteria | Summary of execution |
|---|---|
| **Findability** | (meta)data are assigned globally unique and persistent identifiers which are registered and indexed in searchable resources |
| **Accessibility** | (meta)data are retrievable by their identifier with an open and free protocol, metadata are still accessible even when data is no longer available |
| **Interoperability** | (meta)data use formal, accessible, shared and broadly applicable language and have vocabularies that follow FAIR principles and include qualified references to other (meta)data |
| **Reusability** | (meta)data are associated with accurate and relevant attributes, with detailed provenance, with an accessible license and meet domain-relevant community-standards |

**Table 1: FAIR criteria as described in Wilkinson et al. (2016).**

The FAIR principles are a set of fundamental rules that contribute to good data management and stewardship (long-term care)[29] . The main bottlenecks for putting FAIR in practice for workflows are currently coming from the technical side. For example, workflows are idiosyncratic and encompass a magnitude of different software libraries, software tools and software environments, whose dependencies need to be resolved prior to the actual QC and data analyses steps. Thus, for sharing computational workflows according to the FAIR principles, the software dependencies and scripting code needs to be shared alongside the actual workflow description.

## Annex 2: Software producing and consuming nmrML

To ensure that nmrML will be broadly adopted by life sciences and chemical researchers, we developed webservices and tools covering a large fraction of a typical NMR data acquisition and processing workflow (Figure 1) to generate, convert, process,

[29] Wilkinson et al, 2016: doi:10.1038/sdata.2016.18

validate and publish nmrML files. Additionally, we have worked closely with open source and commercial tool developers to encourage nmrML format adoption. We have summarized efforts already leveraging on the nmrML format in Table 2.

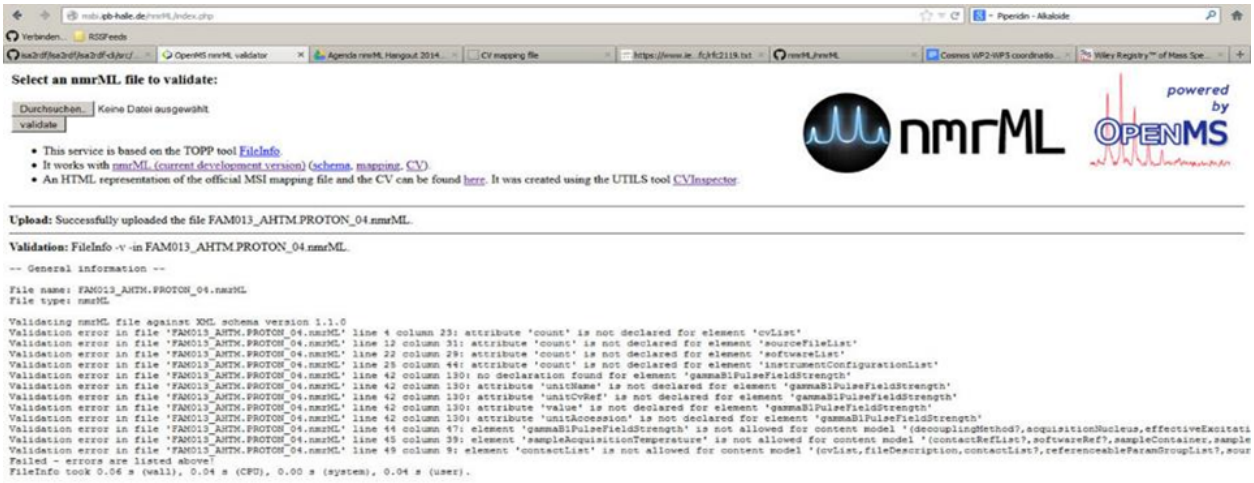
**Table 2: List of nmrML compatible open source software for usage in metabolomics workflows.**

| Tool Category | Tool name | Key Functions | URL | Developer |
|---|---|---|---|---|
| Format Converters | nmrML converter (Java) | Converts vendor to nmrML format | https://github.com/nmrML/nmrML/tree/master/tools/Parser_and_Converters/Java | Institut National de la Recherche Agronomique (INRA), France |
| | nmrML converter (Python) | Converts vendor to nmrML format | https://github.com/nmrML/nmrML/tree/master/tools/Parser_and_Converters/python/pynmrml | The Metabolomics Innovation Center (TMIC), Canada |
| | nmrML to ISA converter | Generate pre-populated ISA files from nmrML files | https://github.com/ISA-tools/nmrml2isa | Ecole Normale Supérieure de Cachan (ENS Cachan), France |
| | BMSxNmrML | Converts BMRB metabolomics entries to nmrML format | http://bmrbdep.pdbj.org/en/bmsxnmrml.html | Institute for Protein Research (IPR), Japan |
| Parsers | MATLAB parser | MATLAB(R) functions parsing and decoding nmrML files, and also writing MATLAB data into nmrML format. | https://github.com/nmrML/nmrML/tree/master/tools/Parser_and_Converters/Matlab | Imperial College London (ICL), United Kingdom |
| | nmRIO | R package for parsing and decoding nmrML files | https://github.com/nmrML/nmrML/tree/master/tools/Parser_and_Converters/R/nmRIO | Leibniz Institute of Plant Biochemistry (IPB), Germany |

| | | | | |
|---|---|---|---|---|
| Data Validators | nmrML semantic validator | XML Schema compliance and rule-based validation of CV usage | http://nmrml.org/validator/ | Leibniz Institute of Plant Biochemistry (IPB), Germany |
| Spectrum Viewers | JSpectraViewer (JSV) | Interactive 1D NMR Spectral viewer used in tools such as Bayesil and nmrML-Assign | http://nmrml.bayesil.ca | The Metabolomics Innovation Center (TMIC), Canada |
| NMR Processing, Identification & Quantification tools | NMRProcFlow | Interactive 1D NMR spectral viewer, spectral processing and quantification tool dedicated to metabolomics | http://nmrprocflow.org | Institut National de la Recherche Agronomique (INRA), France |
| | Bayesil | Automated compound identification, quantification and annotation from 1D NMR spectra | http://bayesil.ca, http://tmic.bayesil.ca | The Metabolomics Innovation Center (TMIC), Canada |
| | nmrML-Assign | nmrML conversion, annotation and peak assignment to compounds for reference 1D NMR spectra | http://nmrml.bayesil.ca | The Metabolomics Innovation Center (TMIC), Canada |
| | Batman | Bayesian deconvolution and automated quantification of metabolites from 1D NMR spectra | http://batman.r-forge.r-project.org | Imperial College London (ICL), United Kingdom |
| | rNMR | Region-of-interest based NMR spectra quantification from 1D and 2D NMR spectra | http://rnmr.nmrfam.wisc.edu | University of Calgary (U of C), Canada |

| | | | | |
|---|---|---|---|---|
| Statistics Tools | MetaboAnalyst | Statistical post-processing | http://www.metaboanalyst.ca | The Metabolomics Innovation Center (TMIC), Canada |
| Workflow Tools | SOMA:tameNMR | NMR data processing and analysis via Galaxy Workflows | https://github.com/pgb-liv/tameNMR | University of Liverpool (UoL), United Kingdom |

## Annex 3: nmrML semantic validator example




Detected errors, as detected by the nmrML validation webservice, are shown here for a particular uploaded nmrML file, which was verified against the nmrML XSD (via the xml validator) and a specific set of checking rules specifying ontology usage (via the OpenMS/Topp semantic validator).

## Annex 4: nmrML Git statistics