Deliverable D2.3

Project Title:	Developing an efficient e-infrastructure, standards and data-flow for metabolomics and its interface to biomedical and life science e-infrastructures in Europe and world-wide
Project Acronym:	COSMOS
Grant agreement no.:	312941
	Research Infrastructures, FP7 Capacities Specific Programme; [INFRA-2011-2.3.2.] “Implementation of common solutions for a cluster of ESFRI infrastructures in the field of "Life sciences"
Deliverable title:	Data exchange format for metabolite quantification
WP No.	2
Lead Beneficiary:	11. IPB
WP Title	Standards Development
Contractual delivery date:	30 September 2013
Actual delivery date:	30. September 2013
WP leader:	Neumann 11. IPB
Contributing partner(s):	1. EMBL-EBI, 11. IPB

Authors: Steffen Neumann, Kenneth Haug

Contents

1 Executive summary

Metabolomics mass spectrometry data analysis usually consists of two major steps 1) feature detection ("peak picking") and grouping into a numerical matrix suitable for 2) statistical analysis such as PCA, uni- or multivariate statistics. The grouping step is not lossless and only retains those features that were detected in multiple samples, and there are no automatic means to conserve metadata such as the experimental design factor assignments. The **mzQuantML** XML data standard was developed by the Proteomics Standards Initiative (PSI) to capture both the individual peak lists and the grouped data matrix for proteomics experiments, while **mzTab** is a standardized spreadsheet like format. The **aim of this deliverable** is to evaluate these data exchange standards and add any required features to fully capture quantification data from Metabolomics experiments.

2 Project objectives

With this deliverable, the project has contributed the following objective:

No.	Objective	Yes	No
1	Develop and maintain exchange formats for raw data and processed information (identification, quantification), building on experience from standards development within the Proteomics Standards Initiative (PSI).	X	

3 Detailed report on the deliverable

3.1 Background

In metabolomics, several community developed software packages exist to process metabolomics data into rectangular data matrices for downstream analysis, such as mzMine2 or XCMS, which have limited integrated statistics support and export the data matrix only into a simple CSV format. The Proteomics Standards initiative (PSI) has developed a number of XML based data exchange standards, e.g. the **mzQuantML** format to store the systematic description of workflows quantifying primarily peptides and proteins by mass spectrometry.

T

his deliverable aims to pave the way for metabolomics software packages to export into the mzQuantML format, which is much more powerful compared with existing tabular representations.

3.2 Description of Work

We had to evaluate the applicability of mzQuantML to metabolomics data (considering small molecules rather than proteins). We created example files and developed prototype parsers for mzQuantML export from the open access tool XCMS.

3.2.1 Analysis of mzQuantML

In a first step, we have analyzed the structure of the mzQuantML format based on the Specification, Schema and more documentation).

The required and applicable top-level XML nodes are <CvList> as part of the data format, the <AuditCollection> and <AnalysisSummary> to capture metadata about the study and the responsible persons. <InputFiles>, <SoftwareList>, <DataProcessingList>, <AssayList> and <StudyVariableList> cover the information about input files, experimental design factors and the data processing steps. While <FeatureList> contains *all* detected peaks in all files, the <AssayQuantLayer> inside the <SmallMoleculeList> contains the numeric data matrix mentioned above.

Our analysis revealed that a lot of the complexity in mzQuantML is not required in metabolomics, i.e. capturing the relationships between proteins and their many peptides actually detected in MS based assays.

3.2.2 Addition of required controlled vocabulary to PSI-MS ontology

We requested several new terms for metabolomics software, which got included in version 3.55.0 of psi-ms.obo ontology. These are required to annotate which software performed the preprocessing and finally exported the preprocessed data into mzQuantML.

3.2.3 Implementation of mzQuantML support in metabolomics software

The first metabolomics MS data processing software to support mzQuantML is XCMS. In addition to the export functions, the unit tests ensure that the produced XML validates against the mzQuantML schema in the version 1.0.0. The required export functionality was added to XCMS in version 1.37.6. This XCMS version is part of the Bioconductor version 2.13 that was released on 15th of October 2013. The export was announced on both the XCMS and PSI mailing lists.

As an example data set we used the Metabolights study MTBLS2, for which we have a full mzQuantML export at https://github.com/sneumann/mtbls2/blob/master/writemzq.mzq.xml

Another software that currently supports mzQuantML is OpenMS. We also contacted other software developers working for and with OpenSource metabolomics software. Nils Hoffmann, main developer of the MaltCMS software also expressed his intention to implement mzQuantML support in MaltCMS. We also contacted Thomas Pluskal, the main developer of the mzMine2 framework. He'll revisit the topic, but currently does not have the required resources.

3.2.4 Overview of available software for downstream analysis

The broad adoption of the MS quantification data standard depends on the availability of parsers for both the data producing software (i.e. data processing packages) and in the software that visualizes and analyzes the data. We here list the available mzQuantML-aware tools.

- **mzQuantML validator**: https://code.google.com/p/mzquantml-validator/
  We contacted the maintainers, and opened several issues (#4 and #5) to fix outstanding problems
- **Generic mzTab conversion**: The **OpenMS** team drafted an experimental support for a conversion from mzQuantML into mzTab, but the support is not yet available.
- **Proteo-Suite** (http://www.proteosuite.org/) is supposed to import mzQuantML files, but crashed with our example, most likely because the software has a hardcoded expectation for protein sequences in the mzQuantML. We opened Issue 19 to inform the developers about this.

The main Issue seems to be that many of the tools are expecting proteomics data, and will fail if e.g. the peptide information is missing. COSMOS will need to continue to stay in touch with the proteomics developer community and provide test data to help these tools to cover a broader range of use cases.

3.2.5 Communication with community

We have contacted the XCMS developer list and the PSI-MS working group, and presented the examples and the implemented XCMS export. We also contacted the mzMine2 core developer, and presented ideas and examples.

3.3 Next steps

With this deliverable we have described the results of evaluation and demonstration to use the PSI-MS standards mzQuantML and mzTab to capture Metabolomics data sets. We'll need to continue to evangelize the use and adoption of the standards, by setting examples and demonstrating the benefits.

4 Publications

None.

5 Delivery and schedule

The delivery is delayed: □ Yes ⍰ No

6 Adjustments made

We have reduced the estimated indicative person months in favour of deliverable D2.4.

7 Efforts for this deliverable

Institute	Person-months (PM)		Period
	actual	estimated	
1. EMBL-EBI	0.5		12
11. IPB	1 (+2 in kind contribution)		12
Total	3.5	16	

Background information

This deliverable relates to WP2; background information on this WP as originally indicated in the description of work (DoW) is included below.

WP2 Title: Standards Development

Lead: Steffen Neumann, IPB

Participants: EBI-EMBL, LU-NMC, MRC, IMPERIAL, TNO and VTT

This work package will deliver the exchange formats and terminological artifacts needed to describe, exchange and query both the metabolomics data and the contextual information (‘experimental metadata’ — e.g., provenance of study materials, technology and measurement types, sample-to-data relationships). We will ensure that these standards are widely accepted and used by involving all major global players in the development process. The consortium represented by COSMOS already contains the majority of players in Metabolomics in Europe and other global players in the field have provided letters of support. Those and others will be invited both the work meetings as well as the regular stakeholder meetings. As the open standards developed here are supported by open source tools, they can be easily put to work which will aid adoption.

Work package number	**WP2**	**Start date or starting event:**									Month 1			
Work package title		Standards Development												
Activity Type		COORD												
Participant number	1: EMBL/EBI	2: LU/NMC	3:MRC	4: Imperial	5: TNO	6: VTT	7:UB	8:MPG	9:UNIMAN	10:CIRMMP	11:IPB	12:UB2	13:UBHAM	14:UOXF
Person-months per participant	12	4	2	3	1	4	2	6	2	6	16	6	4	6

Objectives

1. We will develop and maintain exchange formats for raw data and processed information (identification, quantification), building on experience from standards development within the Proteomics Standards Initiative (PSI). We will develop the missing open standard NMR Markup Language (NMR-ML) for capturing and disseminating Nuclear Magnetic Resonance spectroscopy data in metabolomics. This is urgently needed as long-term archival format if metabolomic databases are to capture all the formats of metabolomic data, as well as supporting developments in cheminformatics and structural biology. For mass spectrometry, we will work with the PSI to extend existing exchange standards to technologies used in metabolomics, e.g. gas chromatography, imaging mass spectrometry and the identification tools and databases.
2. In addition to the raw data formats, we will need to continue the development of standards for experimental metadata and results, independent of the analytical technologies. We will review, maintain and, where needed, extend reporting requirements and terminological artefacts developed by Metabolomics Standards Initiative (MSI). We need to represent quantification options in MS and NMR, and the semantics of data matrices used to summarize experimental results, key information which often is only available in PDF tables associated to manuscripts. As research in biomedical and life sciences is increasingly moving towards multi-omics studies, metabolomics must not be an island. The 'Investigation/Study/Assay' ISA-Tab format was developed to represent experimental metadata independently from the assay technology used. We will use ISA-Tab to standardize metabolomics reporting requirements and terminologies through customized configurations.
3. Finally, we will explore semantic web standards that facilitate linked open data (LOD) throughout the biomedical and life science realms, and demonstrate their use for metabolomics data. While the technical standards already exist, we will need to develop the "inventory" of terms and concepts

required to express facts about metabolomics, capturing the data to characterize studies and digital objects in metabolomics to facilitate the data flow in biomedical e-infrastructures.

Description of work and role of participants

Task 1: Development of data exchange formats for Metabolomics data To capture and exchange raw- and processed mass spectrometry data, we will extend existing open standard (such as mzML, mzIdentML and mzQuantML developed by the PSI) to meet the requirements specific to metabolomics experiments. The MPG will add features missing to handle GC/MS, and the IPB work to represent metabolite identification and -quantitation. MRC will work to promote imzML into an MSI approved exchange format for MS based imaging (MALDI, DESI, SIMS). A new data exchange standard is required for the exchange of NMR spectroscopy based metabolomics data. Building on the excellent experience with XML based formats we will develop the NMR-ML format, a corresponding controlled vocabulary and coordinate the implementation of parsers and tools for validation. Instrument vendors and authors of NMR tools and -databases will be invited to the initiative. The IPB will contribute their expertise from mzML, CIRMMP, including the University of Florence as a third party of CIRMMP, EBI, UBHam and MRC are already involved in discussion with David Wishart from HMDB about NMR-ML.

Task 2: Common representation for Minimum Information Standards for Metabolomics In this WP, we will build on the BioSharing and the ISA-Tab efforts to harmonize representation of the metadata recommendations with other -omics communities, and use automated tests to ensure the interoperability of the metadata between the involved data producers, -consumers and -repositories. The EBI, IPB and MRC will be working with the UOXF to create both core and extended configurations (specific to the research discipline and technologies) suitable for metabolomics, in compliance with the annotation manual created in WP4. This will include a component to report stable isotope labelling and its detection by both mass spectrometry and NMR spectroscopy, required by the metabolomics community carrying out fluxomic studies.

Task 3: Enabling the integration of metabolomics data into large e-science infrastructures. The technologies around the Resource Description Framework (RDF) are used to represent and link the information stored in databases by interconnecting them, relying on a strict semantics for distributed data. Several ontologies of terms and concepts exist for the biological and biomedical domain. In this task we will collect and if necessary extend this inventory to describe metabolomics facts with contributions to existing vocabulary efforts. IPB and UOXF will contribute to e.g. the Ontology for Biomedical Investigations (OBI) and PSI-MS to ensure complete coverage of the key areas of metabolomics technology as a community efforts, leveraging existing, proven infrastructures, in a ‘good citizenship’ frame of mind to avoid duplication of effort. To connect different sources of data and knowledge, the “Semantic Web for Health Care and

Life Sciences Interest Group" (HCLSIG) has started work to represent ISA-Tab metadata as RDF, in compliance with the recommendations of the international Linked Data community (http://linkeddata.org), which will allow to expose any ISA-Tab data set to the semantic web. To demonstrate the feasibility, we will create exemplary semantic query endpoints. The EBI, MPG and IPB will augment their MetaboLights, GMD and MassBank databases. We will also jointly create metabolomics-specific guideline documents for semantic annotation, to maximise the interoperability and link ability of e-resources in the biomedical and life sciences.

Data standards will be described by a set of documents, including 1) the description of use cases, architecture design, and the detailed description of the standard 2) the machine readable standard definition, required for the automatic validation of the content expressed in a standard format 3) several example documents covering the use cases and finally 4) one or more reference implementations. These prototype implementations help to 1) identify shortcomings of the standard definition during the design phase that only crop up during the implementation and practical use, and 2) speed up the adoption in the bioinformatics community that develops metabolomics related software.

The standards defining documents will be discussed during regular phone conferences and at the regular meetings, and developed using open and public repositories. Before they are adopted as MSI standards, they will be sent out to the wider community for a public discussion period. In WP4 we will ensure that international societies and journals make recommendations to use the standards defined in WP2.

Deliverables

No.	**Name**	**Due month**
D2.1	Completion of GC-MS for mzML	6
D2.2	Data exchange format for metabolite identification	12
D2.3	Data exchange format for metabolite quantitation	12
D2.4	Definition of NMR-ML Schema, initial MSI-NMR ontology, example files	12
D2.5	Real data, Converters, Validators and Parsers for NMR-ML	24
D2.6	Collection of ISA configurations for metabolomics studies	27
D2.7	Test infrastructure for the validation of ISA datasets	36

D2.8	Guideline document on RDF and SPARQL for metabolomics resources	24
D2.9	Public availability of query endpoints for linked data from EBI, MPG, IPB	36