Deliverable D2.1

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:	Completion of GC-MS for mzML	
WP No.	2	
Lead Beneficiary:	8. MPG	
WP Title	Standards Development	
Contractual delivery date:	01 04 2013	
Actual delivery date:	01 04 2013	
WP leader:	Steffen Neumann	IPB
Contributing partner(s):	Jan Hummel, MPG and Steffen Neumann IPB	

Authors: Authors: Jan Hummel, Steffen Neumann

Contents

1 Executive summary

Today, most GC-MS data is available either in non-open vendor formats or netCDF. Although netCDF is an open format, it cannot capture for all emerging hyphenated and combinatorial experiment setups, in particular advanced GC-MS experiments, such as Tandem-MS. The aim of this deliverable is to identify and address the limitations, which have so far slowed down the mzML adoption in metabolomics.

2 Project objectives

With this deliverable, the project has reached or the deliverable has contributed to the following objectives:

No.	Objective	Yes	No
1	We will work with the PSI to extend existing exchange standards to technologies used in metabolomics, e.g. gas chromatography	X	

3 Detailed report on the deliverable

3.1 Background

The Proteomics Standards initiative (PSI) has developed a number of XML based data exchange standards. The mzML standard can encode mass spectrometry (MS) raw data, and is widely in use in LC-MS based proteomics, and also increasingly in LC-MS based metabolomics.

However, in GC-MS based metabolomics experiments, data is so far often available as either a closed vendor format or as netCDF (also referred to as ANDIMS), which provides few metadata about the acquisition parameters, and which is unable to capture advanced mass spectrometric experiments such as tandem-MS, which is becoming increasingly popular.

We have thus set out to augment the existing mzML and especially the underlying PSI-MS ontology of controlled vocabulary with terms and concepts required to capture GC-MS based metabolomics experiments, and to further the adoption on both the data producing side and the data processing software and projects

3.2 Description of Work

3.2.1 Collection of a diverse set of GC-MS data files "in the wild"

We have collected a range of GC-MS example data files from both the COSMOS project partners and external contributors. The collected data formats range from vendor files, netCDF to several mzXML and only few mzML files, and have been made available at http://sourceforge.net/projects/cosmos-fp7/.

The aim was to provide a broad range of data and Use Cases that need to be covered by the mzML standard before adoption can be recommended.

3.2.2 *Possible paths to generate mzML data from GC-MS data*

Because mzML is not yet prevalent in the GC-MS based metabolomics world, there is little knowledge in the community how mzML files can be created. We collected the following possibilities:

- **Agilent**: For data in the vendor's MassHunter format, a conversion to mzML is possible with the open source proteowizard (pwiz) software
- **LECO**: the newest version of the vendor software can export mzML. Currently, this software version is not yet common among users. Pwiz is currently not able to convert LECO data, and the company has currently no plans to develop a standalone converter.

- **Bruker**: The company offers several GC-APCI-TOF/MS based instruments. The raw data can readily be converted with both the vendor's CompassXport tool and the pwiz converter.
- **Waters**: Data from e.g. the GCT premier can be converted with the pwiz converter.
- **Thermo**: Data from e.g. the Trace-GC can be converted with the pwiz converter

In some cases the metadata such as acquisition parameters do not exceed that contained in netCDF files.

3.2.3 Identification of existing and missing PSI-MS ontology concepts applicable to GC-MS mzML data, and submission of new ontology concepts to PSI-MS

The mzML data standard uses controlled vocabulary terms from the PSI-MS ontology for specific information to keep the structure of the mzML format stable. We have created and submitted a number of concepts and terms to the PSI-MS ontology working group in the required OBO ontology format which are required to achieve or improve the annotation of MS data.

Based on the collected raw data, we found that only a single vendor is adding additional metadata relevant to GC-MS in netCDF. Here is an example of information present in a LECO netCDF file, and the corresponding existing and recently proposed PSI-MS terms in brackets, if applicable.

```
:test_separation_type = "Gas-Liquid Chromatography" NA;
       :test_ms_inlet = "Capillary Direct" [MS:1000056];
       :test_ms_inlet_temperature = 250.f [MS:1002040];
       :test_ionization_mode = "Electron Impact" [MS:1000389];
       :test_ionization_polarity = "Positive Polarity" [MS:1000130];
       :test_source_temperature = 250.f [MS:1002041];
       :test_accelerating_potential = -600.f [MS:1000304];
       :test_detector_type = "Electron Multiplier" [MS:1000253] ;
       :test_detector_potential = -1850.f [PROPOSED FOR ADDITION] ;
       :test_resolution_type = "Constant Resolution" [MS:1000088];
       :test_scan_function = "Mass Scan" NA;
       :test_scan_direction = "Up" [MS:1000093];
       :test_scan_law = "Linear" [MS:1000095];
       :test_scan_time = 0.0002f [MS:1000502];
```

To obtain a set of “official” vocabulary for the GC-MS instrument models, we contacted all major vendors (LECO, Waters, Thermo, Agilent and Bruker), to provide us with lists and definitions of their products. Those who already responded have been converted to the ontology format used by the PSI, and have also been submitted to the PSI-MS group for inclusion.

Together with the proposed terms, all of this information can now be captured with mzML.

In addition, MPG has extracted a number of concepts and terms, which are used in the Golm Metabolome Database (GMD). They were collected in the COSMOS tracking system at Sourceforge (http://sourceforge.net/p/cosmos-fp7/tickets/), and discussed among the project partners.

3.2.4 Communication with the Bioinformatics and GC-MS development communities to improve mzML adoption

Currently there is a chicken-and-egg problem in the adoption of mzML in metabolomics: few experimentalists convert to mzML (if that is possible in first place) because there is little mzML support in GC-MS specific software, and software developers have little incentive to add mzML import, because few experimentalists produce mzML data.
We have contacted several maintainers of software packages that are able to process GC-MS data, and offered suggestions how to implement the mzML import.
A key factor is the availability of I/O libraries for the popular software development frameworks, which can then be used by specific data processing projects.
For R / Bioconductor can read mzML via mzR: http://bioconductor.org/packages/release/bioc/html/mzR.html
Packages that are mzML-enabled this way include:

- TargetSearch http://bioconductor.org/packages/2.12/bioc/html/TargetSearch.html
- flagme http://bioconductor.org/packages/2.12/bioc/html/flagme.html
- XCMS http://bioconductor.org/packages/2.12/bioc/html/xcms.html
- MSeasy https://sites.google.com/site/rpackagemseasy/

Software developed in Python can use the pymzML library [10.1093/bioinformatics/bts066] at https://github.com/pymzml/pymzML. We have contacted Sean O'Callaghan, one of the maintainers of PyMS https://code.google.com/p/pyms/ , and he intends to implement mzML in PyMS.

Developers in Java can use jmzML at http://code.google.com/p/jmzml/, while C++ users can use either www.open-ms.de or http://proteowizard.sf.net/ to read mzML files.

3.3 Next steps

With this deliverable we have added the required terms and concepts, such that mzML can capture all information contained in today's netCDF files. With this step completed, we have paved the way for further adoption of the PSI-MS data standard in the GC-MS community.
We will continue to support and monitor the adoption of mzML for GC-MS data.

4 Publications

In progress

5 Delivery and schedule

The delivery is delayed: ☑ Yes ☑ No

The last agreement with Waters Company took place in mid April.

6 Adjustments made

N/A

7 Efforts for this deliverable

Institute	Person-months (PM)		Period

	actual	estimated	
8. MPG	1	3	6
11. IPB	1 (+2 in kind contribution)	3	6
14. UOXF	2 in kind contribution		
Total	6	6	6

Appendices

1. N/A

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