Deliverable D2.2

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 identification
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, 8. MPG, 11. IPB

Contents

1 Executive summary

The results of typical metabolomics experiments are usually a table of quantified identified metabolites or unidentified features. The former need to be specified in a way that the actual identified metabolite can be looked up in a number of metabolite databases. Within this deliverable report we describe the applicability and use of the mzTab and mzIdentML standards, the Metabolite Identification focus group of the Metabolomics Society and the CASMI competition (Critical Assessment of Small Molecule Identification).

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

The results of typical metabolomics experiments are usually a table of quantified identified metabolites or unidentified features. The former need to be specified in a way that the actual identified metabolite can be looked up in a number of metabolite databases. Within this deliverable report we describe the applicability and use of the mzTab and mzIdentML standards, and the CASMI competition (Critical Assessment of Small Molecule Identification).

The Proteomics Standards initiative (PSI) has developed a number of data exchange standards. The **mzTab** format was developed to store the end result of an experiment, including peptides and proteins and their quantification.

The **mzIdentML** format was developed to store the full details of protein

identification, and the applicability to Metabolomics needs to be reviewed and evaluated.

3.2 Description of Work

3.2.1 mzTab data format for the reporting of identified metabolites.

The mzTab data format has been developed by the PSI since April 2011, and captures both the quantification and -- if available -- the identification of the measured analyte (i.e. protein or metabolite). The mzTab format contains several sections, including “MTD - Metadata” which contains key-value pairs and the two Table based “PRH/PRT - Protein” and “PEH/PEP - Peptide” sections. For metabolomics the “SMH/SML Small molecule section” is the most relevant.

Currently, the preliminary mzTab for metabolomics format is supported in development versions of the OpenMS, XCMS and CAMERA software tools. We have contacted the authors of the mzMine2 framework and the Maltcms software to discuss the use of mzTab in metabolomics and support them during the implementation in their software.

The MetaboLights database accepts the quantification and identification of metabolites in a subset of mzTab. Based on the initial version of this input format, a second version of this import format has recently been implemented to better incorporate data from NMR based metabolomics experiments. MetaboLights has developed an mzTab export feature to enable open data exchange for both MS and NMR data. The mzTab files are currently available on the MetaboLights ftp site. Similarly, an mzTab import feature is under final development and test. Both export and import features will be incorporated into the “Metabolite identification/annotation plugin” developed for use in ISAcreator.

3.2.2 Evaluation of the applicability of the mzIdentML data format for metabolomics

The mzIdentML standard is designed in analogy to several other PSI data exchange formats. The aim is to capture the input and output of peptide and protein identifications with common proteomics search engines like Mascot, Sequest or OMSA. We have analyzed the schema and documentation to identify which elements are applicable to Metabolomics.

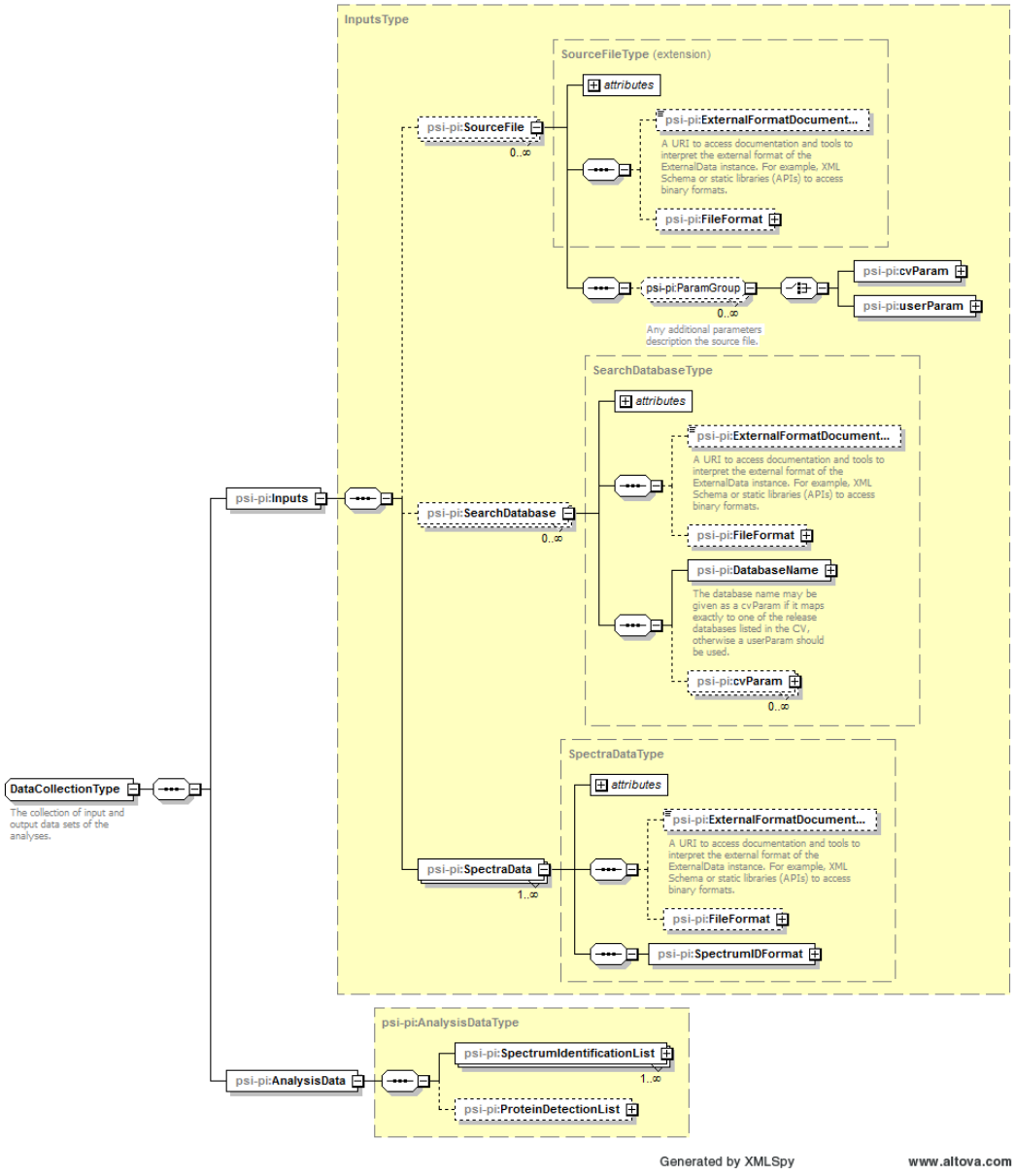

***Figure 1**: Section of mzIdentML to store the actual identification results*

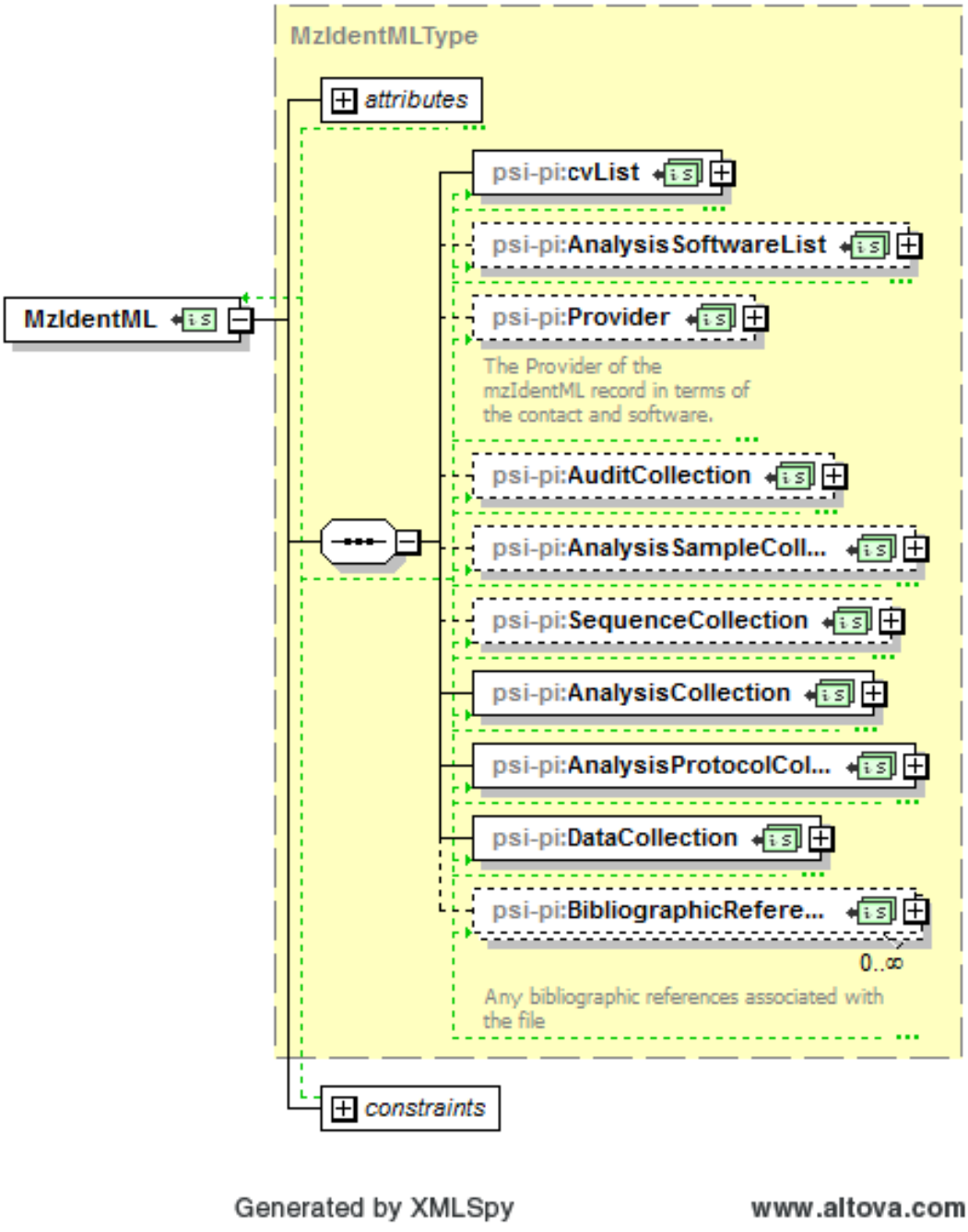

***Figure 2**: mzIdentML section to store the actual protein identification*

Figures 1 and 2 show both the top-level view and a zoom into the results section of the mzIdentML hierarchy from the mzIdentML documentation. Most of the top-level classes in the format are domain unspecific, and can readily be applied to metabolomics experiments.

However, we were unable to find any software reading mzIdentML files that

was not Proteomics specific, and did not find any Metabolomics software that would benefit from the details in an mzIdentML report file. Based on this experience, we are recommending mzTab as the format for reporting metabolite identification in metabolomics experiments.

3.2.3 Metabolite Identification focus group at the Metabolomics Society.

Rick Dunn and Jules Griffin are running the Metabolomics Society Interest Group on Metabolite Identification which also has several other members from COSMOS.

A Metabolite Identification Meeting was held in Manchester in 2012. The meeting was organized by Warwick Dunn with support from "The University of Manchester Investing in Success Funding" scheme, and a significant portion of the COSMOS project members from 1EBI, 2LU/NMC, 8MPG, 9UNIMAN, 11IPB, 13UBHam participated.

The meeting was organized as a round-table discussion and covered several topics, including which platforms are most appropriate for identification, standards for reporting (introduction given by Steffen Neumann), whether MS/MS is adequate or MSn is required, Mass spectral databases and libraries and finally *de-novo* structure elucidation.

3.2.4 Metabolite Identification contest "CASMI".

In 2012/13, the IPB and the eawag institute (Zürich, CH) jointly organized the identification contest CASMI: the *Critical Assessment of Small Molecule Identification*. We published challenge spectra without revealing the identity of the measured analyte, and invited the community to submit identification hypotheses.

After the contest, we published the proceedings of the CASMI contest in a special issue "Small Molecule Identification beyond the Crystal Ball - Selected Papers from CASMI" of the Open Access, peer-reviewed MDPI journal Metabolites. We also presented a poster on CASMI at the 61st ASMS Conference on Mass Spectrometry and Allied Topics June 9 - 13, 2013 in Minneapolis.

The contest allowed an unbiased overview of several metabolite identification approaches, and the collection of the submission metadata will be used in the refinement of the mzTab reporting recommendations.

In 2013/14 the contest is organized by a team of Japanese experts around Prof. Takaaki Nishioka (Nara Institute of Science and Technology), and initial contacts have been made with the putative organizers of CASMI 2014/15, so it will continue to serve as an evaluation contest and possible test-bed for metabolite identification reporting.

3.3 Next steps

We scheduled a workshop on the topic of "mzTab for metabolomics" together with Oliver Kohlbacher from Tübingen University, who is one of the leading members in the mzTab development in the proteomics standards initiative PSI. The workshop will take place on March 6th 2014 at the University of Tübingen.

4 Publications

Special Issue "Small Molecule Identification beyond the Crystal Ball - Selected Papers from CASMI", Metabolites (ISSN 2218-1989).

Schymanski E.L. & Neumann, S. The Critical Assessment of Small Molecule Identification (CASMI): Challenges and Solutions. *Metabolites* **3(3)**, 517-538, (2013)

Neumann, S., Rasche, F., Wolf, S. & Böcker, S. Metabolite Identification and Computational Mass Spectrometry. In: *The Handbook of Plant Metabolomics, Metabolite Profiling and Networking* (Weckwerth, W. & Kahl, G.). *Wiley-VCH* 271-285, (2013) ISBN: 978-3-527-32777-5

Schymanski E.L. & Neumann, S. CASMI: And the Winner is… *Metabolites*

3 (2), 412-439, (2013)

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
8. MPG	1		12
11. IPB	1 (+2 in kind contribution)		12
Total	4.5	14	

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