Deliverable 3.2

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:	Integrable technology-specific software tools
WP No.	3
Lead Beneficiary:	8. MPG
WP Title	Database Management System
Contractual delivery date:	1 10 2014
Actual delivery date:	1 10 2014
WP leader:	Dirk Walther 3. MPG
Contributing partner(s):	3. MPG, 4. Imperial college London, 7. University of Barcelona, 8. MPG, 10. Florence University, 11. IPB, 12. University Bordeaux, 14. UOXF, External partner/ Stakeholder D. Wishart, University of Alberta,

Authors: Dork Walter

Contents

1 Executive summary

A set of 20 computational tools and services comprising small format conversion facilities to advanced spectra interpretation software packages that address unmet needs in metabolomics data processing have been developed by the COSMOS consortium partners and made available to the Metabolomics community. Available at COSMOS website (http://cosmos-fp7.eu/tools)

2 Project objectives

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

No.	Objective	Yes	No
1	Integrable technology-specific software tools (as web services or Galaxy-compliant software components etc.), M24	X	

3 Detailed report on the deliverable

3.1 Background

Reliable and reproducible data analysis hinges on standardized data processing. Ideally, isolatable steps along the processing pipelines are identified and software solutions developed for them that are then strung together into consistent workflow that can be easily applied by anyone confronted with similar tasks.

3.2 Description of Work

COSMOS partners identified unmet needs for software support facilities and developed tools and services to address them. As the partners employ a broad range of technologies, the created toolset is equally diverse and ranges from little converter tools to advanced statistical spectra interpretation software packages. A detailed overview of the developed tools is provided below.

3.2.1 Developed tools and services

Summary for the tools developed by COSMOS partner are:

	Name of Service	Developer (COSMOS partner)	URL/ Codebase	Description
1	BATMAN	4:Imperial College London	http://batman.r-forge.r-project.org/	BATMAN is an R package for estimating metabolite levels in Nuclear Magnetic Resonance spectral data using a specialised MCMC algorithm. It deconvolves peaks from 1-dimensional NMR spectra, automatically assigns them to specific metabolites from a target list and obtains concentration estimates. The Bayesian model incorporates information on characteristic peak patterns of metabolites and is able to account for shifts in the position of peaks commonly seen in NMR spectra of biological samples.

2	Metassimulo	4:Imperial College London	http://cisbic.bioinformatics.ic.ac.uk/metassimulo/	MetAssimulo is a MATLAB-based package which simulates 1H-NMR spectra of complex mixtures such as metabolic profiles. Drawing data from a metabolite standard spectral database in conjunction with concentration information input by the user or constructed automatically from the Human Metabolome Database, MetAssimulo is able to create realistic metabolic profiles containing large numbers of metabolites with a range of user-defined properties
3	MZmine_2		http://mzmine.sourceforge.net	MZmine 2 is an open-source project delivering a software for mass-spectrometry data processing, with the main focus on LC-MS data. It is based on the original MZmine toolbox described in 2006.
4	MIDcor	7:U Barcelona	http://sourceforge.net/projects/gcmscorrection/files/?source=navbar	MIDcor (Mass Isotopomer Data corrector) is an R-based computer program designed for the extraction of “pure” 13C mass isotopomer distribution of metabolites formed from artificial 13C-enriched substrates. In addition to subtraction of natural isotope distribution from raw m/z data it examines a possible overlapping m/z peaks for several metabolites, and corrects it if such an overlapping takes place.
5	SpecView	8:MPIMP/ MPG	http://gmd.mpimp-golm.mpg.de/download/	SpecView 1.0 is an Microsoft Excel™ plugin that facilitates the presentation of deconvoluted GC-MS Spectra in Microsoft Excel. Selected MS-Spectra (format: '87:100 103:78') are normalized prior to drawing.
6	KODAMA	10:CIRMMP, Florence	http://www.kodama-project.com/download.html	KODAMA is an innovative method to extract new knowledge from noisy and high-dimensional data, and offers a general framework for analyzing any kind of complex data in a broad range of sciences. It is particularly suited, but not limited, to cluster metabolomic data. Ref.: http://www.pnas.org/content/111/14/5117.abstract

7	mzR	11:IPB	https://github.com/sneumann/mzR/	Framework for processing and visualization of chromatographically separated and single-spectra mass spectral data. Imports from AIA/ANDI NetCDF, mzXML, mzData and mzML files. Preprocesses data for high-throughput, untargeted analyte profiling.
8	xcms	11:IPB	https://github.com/sneumann/xcms/	Framework for processing and visualization of chromatographically separated and single-spectra mass spectral data. Imports from AIA/ANDI NetCDF, mzXML, mzData and mzML files. Preprocesses data for high-throughput, untargeted analyte profiling.
9	CAMERA	11:IPB	https://github.com/sneumann/CAMERA/	Annotation of peaklists generated by xcms, rule based annotation of isotopes and adducts, EIC correlation based tagging of unknown adducts and fragments
10	Rdisop	11:IPB	https://github.com/sneumann/Rdisop	Identification of metabolites using high precision mass spectrometry. MS Peaks are used to derive a ranked list of sum formulae, alternatively for a given sum formula the theoretical isotope distribution can be calculated to search in MS peak lists.
11	nmrML validator	11:IPB	http://nmrml.org/validator/	This service is based on the TOPP tool FileInfo. It works with nmrML using the current development versions of the schema, mapping and CV. An HTML representation of the official MSI mapping file and the CV can be found online.
12	metfRag	11:IPB	https://github.com/c-ruttkies/MetFragR	Identification of metabolites using high precision mass spectrometry. Candidate molecules of different databases are fragmented in silico and matched against mass to charge values. A score calculated using the fragment peak matches gives hints to the quality of the candidate spectrum assignment.
13	nmRIO	11:IPB	https://github.com/sneumann/NMR-ML/tree/master/tools/Parser_and_Converters/R/nmRIO	Parser for NMR raw fid and processed data from nmrML, Bruker and Varian/Agilent

14	nmrML java converter	12:U Bordeaux	http://nmrml.org/converter/	Based on both nmrML.xsd (XML Schema Definition) and CV params (such as ontologies nmrCV, UO, CHEBI ...), a converter written in Java was developed that automatically generates nmrML files, from raw files of the major NMR vendors
15	Xeml interactive Designer - (Xeml-Lab)	12:U Bordeaux	https://github.com/cbib/XEML-Lab	Xeml-Lab 1.0 is a software initially developed at the Max Planck Institute in 2008-2009, in order to provide an interactive graphical interface, allowing the design of complex experiments, and the generation of machine-readable metadata files. The main goal of the XEML project was to realize a standard for the control and documentation of experimental design and growth, with a maximum of power in terms of data processing, not only to obtain reliable molecular and physiological data, but also to make plant growth metadata amenable for studies in integrative biology. To make Xeml-Lab, easily distributable on various platforms (Windows, Mac and Linux), and to offer a wider access to this tool to the biological community, we decided to rewrite it in C++ language. The new version Called Xeml-Lab 1.0 is now available.
16	rISA	14:UOXF	https://github.com/ISA-tools/Risa	The Investigation / Study / Assay (ISA) tab-delimited format is a general purpose framework with which to collect and communicate complex metadata (i.e. sample characteristics, technologies used, type of measurements made) from experiments employing a combination of technologies, spanning from traditional approaches to high-throughput techniques. Risa allows to access metadata/data in ISA-Tab format and build Bioconductor data structures. Currently, data generated from microarray, flow cytometry and metabolomics-based (i.e. mass spectrometry) assays are supported.

17	LinkedISA	14:UOXF	https://github.com/ISA-tools/linkedISA	linkedISA is a package to convert ISA-TAB files into RDF data relying on different semantic frameworks (OWL ontologies). For more information, please visit the linkedISA website: http://isa-tools.github.io/linkedISA/.
18	ISATab Viewer	14:UOXF	https://github.com/ISA-tools/ISATab-Viewer	Render from files on your web server: Provided the files are hosted on the same server your page is running on, you can just do this to load in your ISA files. Now integrated with BMC GigaScience and allows out of box rendering of ISA-Tab formatted archives for consistent and easy content exposure of experimental metadata.
19	Biocrates2ISA.xsl	14:UOXF	https://github.com/ISA-tools/xsl2isa	an XSL transformation allowing direct deposition from Biocrates AG software to Metabolights or publication of targeted Metabolomics data generated on the Biocrates Platform as ISA-Tab metadata.
20	nmrML python converter	external partner/stakeholder: U Alberta	https://github.com/nmrML/nmrML/tree/master/tools/Parser_and_Converters/python/pynmrml	pynmrML is a Python library for reading, writing and interfacing with nmrML documents. It includes nmrML bindings, nmrCV bindings, and readers and writers for several different NMR formats. It relies on the excellent NMRglue library for processing raw NMR data.

3.2.2 Communication with community

Steffen Neumann attended ASMS 2013, where a Galaxy workshop was organised by the proteomics community, and discussed the opportunities and requirements from a metabolomics point of view.

3.3 Next steps

We will explore and pursue opportunities to integrate the created software tools into workflow systems such as provided by the Galaxy.

The Galaxy environment provides user-friendly interface (e.g. a workflow editor) as well as convenient functionalities to store and share workflows (modules and their parameters) and histories (input and output data from each analysis).

Modules and associated documentation enable users to build their own workflow, run analyzes, visualize and download the results. Modules necessary to build a full LC-MS workflow (pre-processing with XCMS and CAMERA, analytical drift correction, uni- and multivariate analysis, and annotation with public databases) are already available.

4 Publications

None.

5 Delivery and schedule

The delivery is delayed: ☐Yes ☑No

6 Adjustments made

N/A

7 Efforts for this deliverable

Institute	Person-months (PM)	
	actual	estimated
1: EMBL-EBI	2	
5: TNO	0.74	
4: Imperial	1.12	
6: VTT	0.47	
7: UB	3	
8: MPG	4	
2: LU	2	
11: IPB	2	
12: UMAN	1	
14: UOXF	2.16	
Total	**17.99**	**24**

Appendices

1. N/A

Background information

This deliverable relates to WP3; background information on this WP as originally indicated in the description of work (DoW) is included below.

WP3 Title: Database Management System

Lead: Dirk Walther, MPI Molecular Plant Physiology, Golm (MPG)
Participants: EBI-EMBL, LU-NMC, MRC, IMPERIAL, TNO, VTT, UB, MPG, UNIMAN, IPB, UB2, UBHam, UOXF

This work package will focus on developing and coordinating the infrastructure to easily access, to process, store, and exchange metabolomics measurement and associated experimental metadata.

Work package number	WP3	**Start date or starting event:**						Month 1						
Work package title	Database Management System													
Activity Type	COORD													
Participant number	1: EMBL-EBI	2: LU/NC	3:MRC	4:IMPERIAL	5:TNO	6:VTT	7:UB	8:MPG	9:UMAN	10:CIRMMP	11:IPB	12:UB2	13:UBHAM	14:UOXF
Person-months per participant	9	4	2	3	1	4	7	14	2	0	6	6	2	4

Objectives

This work package will focus on developing and coordinating the infrastructure to easily accessible to[d1] process, store, and exchange metabolomics measurement and associated experimental metadata. Specifically, four central development areas will be worked on:

1) Capturing and exchanging experimental metadata,
2) Technology-specific data handling and processing,
3) Management and integration of generalizable metabolomics data, and

1. 4) Integration of metabolomics data with all other levels of molecular organization

Description of work and role of participants

It is in the very nature of a coordination action to focus on communication between the participants for the sake of policy making, to document the outcome and spread the word to promote widespread community adoption.
We therefore wish to highlight the following:

Task 1, Experimental Metadata: We will extend relevant components of the ISA software suite and establish it as de-facto standard for experimental metadata deposition. The UOX team behind the ISA framework will head this effort. If and where needs are identified, the ISA-Tab syntax specifications will be further developed to make complex study design exchange possible. ISA-Tab is already implemented by several tools and used by a growing numbers of communities in several life science domains. It is pivotal that any extension does not compromise the current structure and backward compatibility is addressed. All partners will give input during this phase and any extension will be also presented to the existing ISA user community, as these may have an impact on them and certainly will have an impact on the ISA software suite. Coordinated action aiming at enabling exchange of study data between the NuGO (NutriGenOmics) phenotype database, the data support platform of LU/NMC (Partner 2) and EBI MetaboLights (Partner 1) will be enacted. Data flow between the aforementioned centers can be greatly facilitated by the creation of shared curation practice, leading to the creation of a pool of experts within the coordination action. Their practice will lead to the development of guidelines to consistently describe common patterns of experimental design and UOXF team will lead on this activity. The ultimate goal for metadata handling will be to standardize the “feed-in” data flows of meta-information into the centralized European metabolomics database MetaboLights, see task 3.

Task 2: Technology-specific processing and handling software: Metabolomics technologies are diverse and require specialized software infrastructure, processing and analysis tools. Towards unifying the software solutions we will identify common design principles, data formats for efficient data exchange and comparison. The MPIMP will continue its focus on the GC/MS technologies. University of Barcelona will concentrate on the software for processing and handling 13C tracer metabolomics data. Such software will be developed based on the tool Isodyn (from “isotopomer dynamics”), already developed by this team. It will be adopted for processing the 13C distribution data obtained with GC/MS as well as NMR technologies, and accepting the formats in which they are presented in the existing databases. The result of such data processing will be the distribution of metabolic fluxes corresponding to the analysed distribution of isotopic isomers. The software will be adopted to store the results of analysis in available databases. MRC (Partner 3) and IPB will examine software platforms for LC-MS and other mass spectrometry based (such as MALDI, DART, etc) approaches for both aqueous metabolites and lipidomics. MRC and University of Manchester (Partner 9) will follow the development of imaging technologies for mass spectrometry - a novel area that holds the potential of applying MS-based analysis to histology studies of tissues. UB2 (partner-13) will concentrate on the NMR data. At the level of raw-data, the individual technology-centric software platforms will develop data storage and handling policies and protocols to guarantee persistent and safe data storage. With regard to distributing specialized software solutions and making them easily accessible to the scientific community, we intend to promote the use of and build on web-service software solutions and/ or workflow components that are seamlessly integrable in custom or standardized processing workflows (e.g. using the Galaxy workflow management system).

Task 3, Developing MetaboLights as the centralized metabolomics data hub As a repository for higher-level metabolomics data; i.e. summarized, processed data, all partners recognize the MetaboLights as the central integration hub. Further developing MetaboLights will be the task of the EBI (Partner 1, Christoph Steinbeck and team.) Specifically, every technology-specific dataset needs to be rendered MetaboLights compatible. Syntax standards, such as ISA-Tab need to be promoted and supported to facilitate optimal exploitation of experimental meta-information for dataset discovery. In addition, The format also needs refinement to ensure data integration at the level of processed metabolite data. This requires refining the data files -associated to the ISA-Tab- that report of metabolite identifications, quantification in individual samples but also consistent reporting of group comparisons and to be able to do so in a range of specific applications for monitoring metabolites associated to chemical families and associated analytical techniques. Finally, data need to be deposited to road test the validity of the meta data descriptions. MRC (Partner 3), the IPB (Partner 11), the MPG (Partner 8) and (Partner 12) will deposit datasets from NMR spectroscopy, GC-MS and LC-MS studies. These datasets will also provide important resources for others to develop software solutions to the metabolomics pipeline.

Task 4, Integration with other levels of molecular organization: Integration of metabolomics data with data from other domains of molecular organization such as genomics, transcriptomics, and proteomics has been recognized as critical for a meaningful interpretation of metabolomics data. Towards facilitating the data integration, we will establish annotation standards (together with WP2), ID-mapping routines, and further develop reporting standards to easily integrate metabolite data with metabolic pathway information (KEGG, Biocyc/Pathway tools software). Link-out capacities will be primarily offered via the MetaboLights resource (Partner 1). Biological relevant questions will be used for the development of the links, TNO (Partner 5) will be responsible for this part. To road-test these links ‘poly-omic’ studies using metabolomics alongside transcriptomics and proteomics will be deposited by MRC (Partner 3) and TNO (Partner 5). Existing datasets, such as the InnoMed PredTox available from the BioInvestigation Index (BII) instance at EBI already demonstrating how ISA-Tab syntax has been used to encode such complex experiments, will be reassessed and analysed in order to evaluate how improvements can be made or lessons learned from this experience.

Deliverables

No.	Name	Due month
D 3.1	Software infrastructure for capturing and exchanging metadata	18
D3.2	Integrable technology-specific software tools	24
D3.3	Deposition of 50 standardised community datasets in MetaboLights	12