Deliverable D3.1

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:	Software infrastructure for capturing and exchanging metadata and channeling whole data sets into Metabolights (Feed-in standards)	
WP No.	3	
Lead Beneficiary:	8: Max-Planck-Institute of Plant Physiology	
WP Title	Database Management System	
Contractual delivery date:	1 April 2014	
Actual delivery date:	1 April 2014	
WP leader:	Dirk Walther	MPIMP
Contributing partner(s):	Benjamin Dartigues, Macha Nikolski; University of Bordeaux (CBIB)	

Authors:

Kenny Billiau (MPIMP), Jan Hummel (MPIMP), Dirk Walther (MPIMP), Benjamin Dartigues (CBIB), Macha Nikolski, (CBIB),

CONTENTS

1. Executive summary

A software infrastructure was established that enables the convenient capture of standardized experimental metadata. It builds on the existing XEML-lab software suite (Hannemann et al. 2009) that was developed for a graphics-supported description of experimental designs along with standardized description of experiment conditions via ontologies. XEML-labs has been expanded to facilitate data import and export and a seamless integration to existing databases. In particular, the export to ISA-tab format has been added to allow convenient upload to MetaboLights metabolomics data repository. XEML-labs re-implemented for available for all common operating systems. Its integration into an existing database environment has been successfully demonstrated for the Golm Metabolome Database (GMD). With this deliverable, a significant contribution has been made towards broad adoption of standardized description of experimental metadata.

2. Project objectives

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

No.	Objective	Yes	No
	Developing the XEML framework	☑	
	Submitting metabolite profiling experiments into *MetaboLights*	☑	
	Testing the ISA Tools and giving feedback to the *MetaboLights* developers	☑	

3. Detailed report on the deliverable

3.1. Background

In a single-user environment, experimental metadata annotation can be efficiently handled using *ISAcreator*, part of the *ISA-Tools* software suite. As larger institutional environments might already have established proprietary tools for their primary research data and looking for convenient ways to share and integrate those with *MetaboLights, COSMOS* also aims to standardize such workflows. Interfacing with dedicated databases utilising alternative metadata annotation tools will engage and enable a broad user base to export data from their local systems into ISA-Tab formatted data sets, and subsequently to easily import or submit to *MetaboLights*. Alternative tools include the *eXtensible Experiment Markup Language* (XEML) to automate processing pipelines within Bioconductor packages and general *BioPortal* (http://bioportal.bioontology.org/) powered ontologies.

The external resources listed below provide ISA formatted exports:

- *The Golm Metabolome Database* (GMD)
- *The Metabolomic Repository Bordeaux* (MeRyB)
- *The Netherlands Metabolomics Centre Data Support Platform / Phenotype Database* (NMCDSP/dbNP)
- R - Package 'Risa' http://www.bioconductor.org/packages/release/bioc/html/Risa.html

Considering the high diversity and breadth of metabolomics applications and applied analytical technologies, metabolomics applications clearly lack reporting standards for experimental objectives and provenance of the analysed materials. To report specifically on plant metabolomics experiments, the Max Planck Institute

for Molecular Plant Physiology (MPIMP) started with the development of the XEML framework. XEML was conceptualised as a decoupled layer on top of separate, independent and dedicated databases in 2006. As a machine readable XML dialect, the schema based XEML provides means to store experimental design and metadata describing the actual experiment, together with references to one or more independent databases hosting the actual experimental results. Following a strict decoupling of data producers and data consumers by means of predefined software engineering interfaces, specialised tools were conceptualised, capable of compiling metadata about experimental designs and actual data from data providers (e.g. databases) into numerical matrices for statistical evaluation in applications such as statistical environment R, Microsoft Excel™ or MATLAB™.

Although the centralisation of experimental results into an XEML-Store was already conceptualised by Hannemann et al. (2009), this feature had not been implemented yet. Following XEML's general ideas, we identified the need to have a relational data model available for evaluation of actual experimental results in combination with metadata and to provide a store of experiments for data mining purposes. Hence, we contributed a relational data model to the open source XEML framework and further designed a software interface to compile XEML based metadata into ISA-Tab formatted files.

We further developed and improved the XEML graphical designer. Nowadays, typical plant metabolite profiling experiments tend to have hundreds of samples and therefore, in terms of simplicity and scalability, we favour the graph based experiment visualisation over a table based textual description of experimental settings. These implementations can then also provide possible prototype ideas for later ISA-Tool versions.

As the description of the analytical platform is the key for successful experimental comparison of analytical results, we undertook major efforts of implementing an

ontology based annotation of the gas-chromatography mass-spectrometry (GC-MS) variants utilised at the MPIMP. This task included searching for appropriate terms and, where appropriate terms could not be identified or were not available, contacting the ontology's owner to request the new term from the ontology maintainer. The results were subsequently used for the export to the ISA-Tab file format.

We helped to extend and reviewed relevant components of the ISA-Tools software suite and, by doing so, helped to establish the ISA-Tab file format as de-facto standard for experimental metadata deposition in *MetaboLights*.

Based on our use cases, we did not experience any inconsistencies with the ISA-Tab syntax specifications, even when testing with respect to complex study designs.

Starting from plant specific experiments and potentially increasing complexity towards more general designs, we utilised 10 years of experience of the MPIMP, catalogued in GMD's experiment repository to develop and validate guidelines to consistently describe common patterns of experimental design using the ISA-Tools framework.

3.2. Description of Work

Promoting the data integration proposed by COSMOS requires actual data. Instead of relying on assumptions about unspecified data, we decided to drive this process using existing datasets, in particular the Golm Metabolome Database (GMD). Although the accomplished developments are specifically focusing on the GMD, they are helpful and necessary to identify general solutions, patterns and practices in the data-sharing process and to evolve from a specific towards generic implementations.

The GMD is a repository of results from GC-MS based metabolite profiling experiments. To this date, the GMD contains 235 experiments with 31,313 chromatograms. While 11 experiments are publicly accessible, the remaining experiments are proprietary. For data analysis and data evaluation, we integrated XEML based metadata to describe environmental conditions and sampled plant materials in terms of structure and developmental stages into the GMD. Although the GMD is historically focused on plant metabolomics experiments, such specialisation can commonly be expected in other dedicated data repositories as well.

3.2.1. Extending the extensible mark-up language

Based on feedback from users, we developed and maintained the graphical XEML designer software. In particular, we developed and implemented a relational data schema, as further described below. While the MPIMP concentrated on developing a windows version, our collaboration partners at the University of Bordeaux (UB) ported the application to Linux and Macintosh. We successfully validated these different development versions against each other. Part of this collaboration was a staff exchange between UB and MPIMP.

3.2.2. Porting the XEML Designer

One of the challenges that we faced in this work package was to write a new version of XEML Designer to allow its use on different platforms (Mac, Windows, and Linux). This work was the responsibility of CBiB at the University of Bordeaux. We decided to translate the code, originally written in C# into C++. Moreover, a user-friendly and easy to maintain user interface is of great importance if we want the tool to be widely used. For these reasons, we have chosen the Qt as our graphics library. Indeed, one of the nice features in Qt is that it provides a broad

set of libraries in addition to the GUI part (e.g. XML parsing, threads, networking), all in a consistent style and all multi-platform. The flexible Graphic Widget Management was an additional reason of this choice. After translating and porting all necessary classes (about 170 of them) that correctly builds an Xeml file (Figure 1), the result is a multi-platform application that respects the initial Xeml Designer and is capable of displaying every component of an experiment (Story, Event, Observation point) as well as its related metadata within a user-friendly graphical interface.

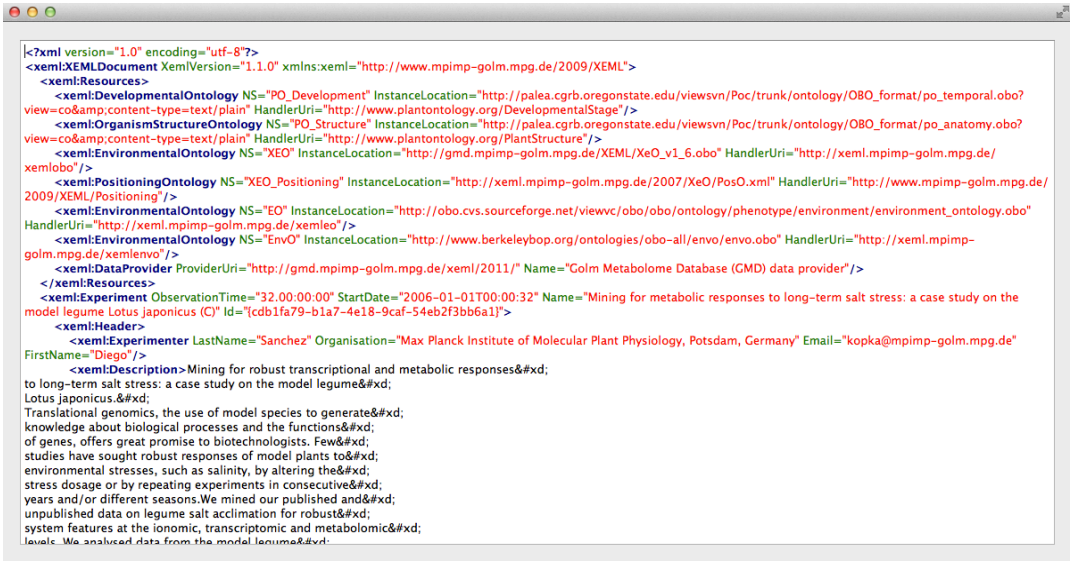
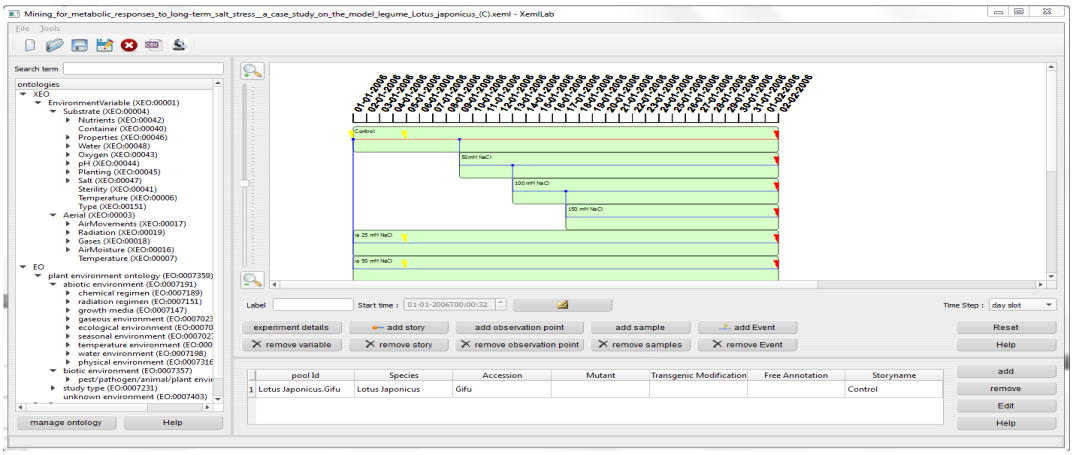


Figure 1: XEML Document Editor.

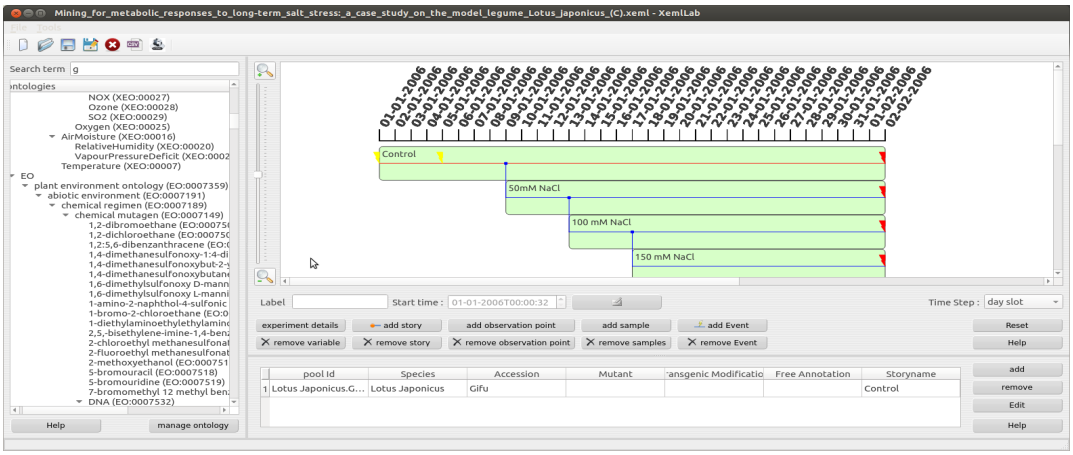
Many new or improved functionalities are embedded within the new XEML Lab designer, including the import of environmental data from a tabular file (csv, tsv format), which was developed at the CBiB in Bordeaux. Some of these novel functionalities are still under development, such as plotting every quantitative variable during the time of experiment or exporting to ISA-tab format using a new relational data model (section 3.2.4). An ability to discriminate set variables versus

measured variables has been added to XEML document, reflecting the fact that some environmental variables can be fixed before the experiment (e.g., substrate type) while others can be measured during the experiment (e.g., T°). After capturing metadata using ontologies available in XEML Designer, it is now possible to describe six types of variables from environmental, structural and developmental stage ontologies (XEO, EnvO, EO, Xeo_positioning, PO_developmental and PO_Structure). Below is a screenshot of XEML Designer's Graphical Interface for each platform (Figure 2 a, b, c). The XEML Designer sources and binaries for each platform can be downloaded from: https://github.com/cbib/XEML-Lab

(a)



(b)



(c)

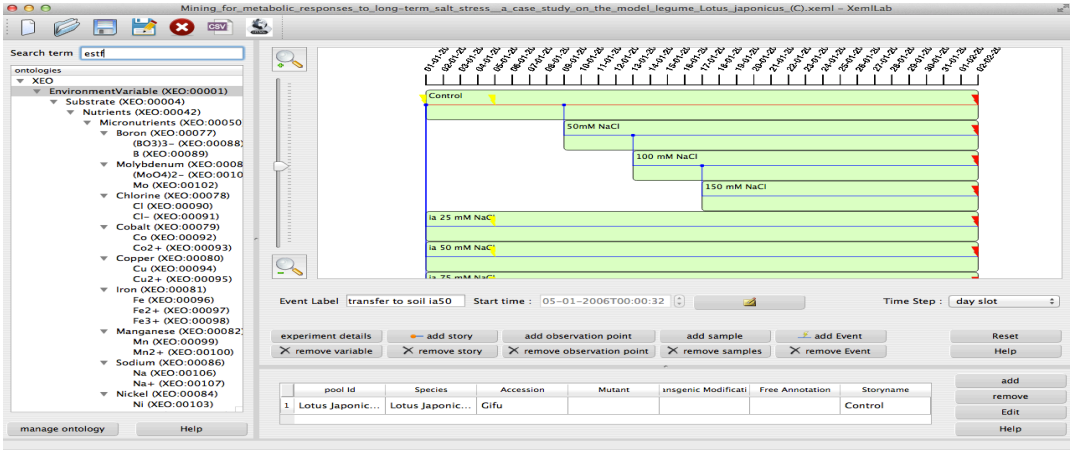


Figure 2: XEML Designer graphical interface on Windows (a), Linux (b) and Mac (c). You can see the Ontologies tree on the left panel, the experiment view on the top right and Genotype view on the Bottom right

3.2.3. Developing a connection to PLATO database

XEML Designer was originally designed to collect environmental metadata from biologists' experiments and better describe these experiments in terms of environmental and structural information of the plants under study. This is what we commonly call « metadata ». However, in order to gain biological insight these

environmental data have to be correlated with experimental data stored in related databases. To illustrate this, a connection with the high-throughput biochemical phenotyping “PLATO” database hosted at the INRA of Bordeaux was implemented. PLATO is a micro plate processing pipeline and supports enzyme activities and metabolites contents. PLATO, was developed by Jan Hannemann at INRA-Bordeaux, and is an evolution of the EnzymeLaborTool developed at the Max-Planck Institute of Molecular Plant Physiology (http://plato.codeplex.com/). PLATO hosts a collection of large data sets of robotized biochemical phenotyping (Figure 3). PLATO connection is fully functional, but users need to have a granted access to the database in order to map their environmental data with their experimental data.

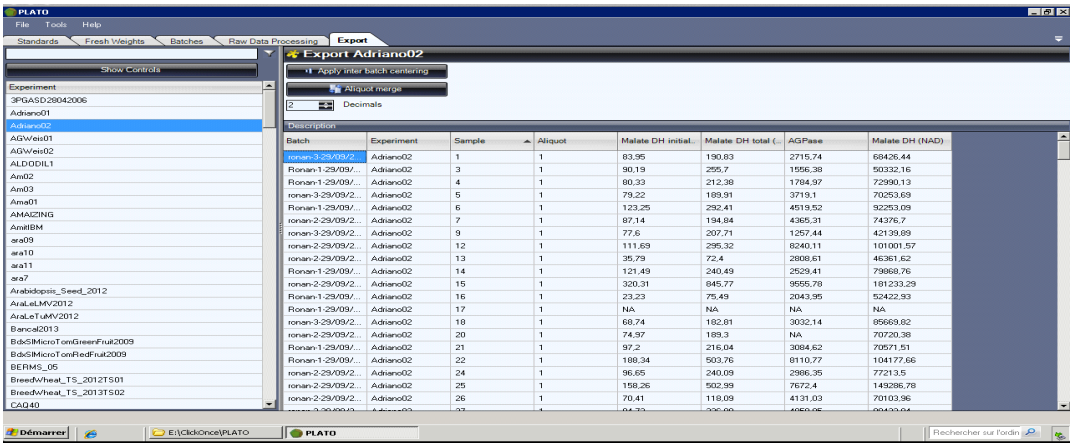
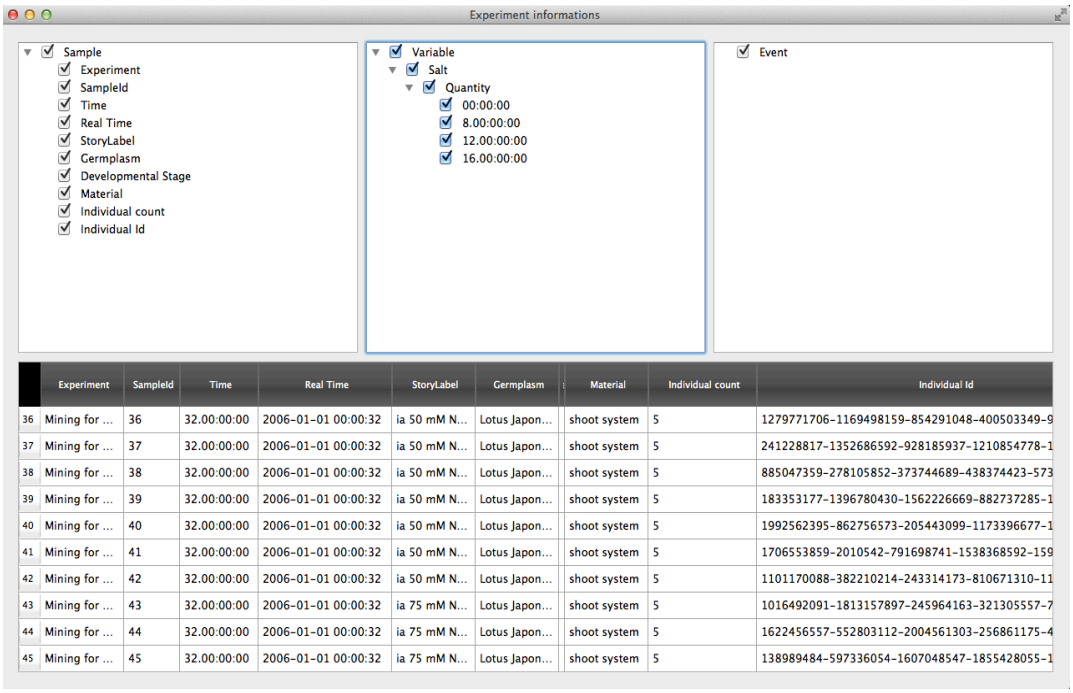


Figure 3: Screen shot of PLATO database

PLATO contains a large number of plant-development and plant physiology-related experiments enabling rich linking of experimental and environmental metadata (see section 3.2.6), thus enables a wide usage of XEML Designer by biologists. All these data can be displayed in the table of experiment, allowing an overview of all experiment and the possible export of all these data to some tabular format for further statistical analysis (Figure 4).



	Experiment	SampleId	Time	Real Time	StoryLabel	Germplasm	Material	Individual count	Individual Id
36	Mining for ...	36	32.00:00:00	2006-01-01 00:00:32	ia 50 mM N...	Lotus Japon...	shoot system	5	1279771706-1169498159-854291048-400503349-9
37	Mining for ...	37	32.00:00:00	2006-01-01 00:00:32	ia 50 mM N...	Lotus Japon...	shoot system	5	241228817-1352686592-928185937-1210854778-1
38	Mining for ...	38	32.00:00:00	2006-01-01 00:00:32	ia 50 mM N...	Lotus Japon...	shoot system	5	885047359-278105852-373744689-438374423-573
39	Mining for ...	39	32.00:00:00	2006-01-01 00:00:32	ia 50 mM N...	Lotus Japon...	shoot system	5	183353177-1396780430-1562226669-882737285-1
40	Mining for ...	40	32.00:00:00	2006-01-01 00:00:32	ia 50 mM N...	Lotus Japon...	shoot system	5	1992562395-862756573-205443099-1173396677-1
41	Mining for ...	41	32.00:00:00	2006-01-01 00:00:32	ia 50 mM N...	Lotus Japon...	shoot system	5	1706553859-2010542-791698741-1538368592-159
42	Mining for ...	42	32.00:00:00	2006-01-01 00:00:32	ia 50 mM N...	Lotus Japon...	shoot system	5	1101170088-382210214-243314173-810671310-11
43	Mining for ...	43	32.00:00:00	2006-01-01 00:00:32	ia 75 mM N...	Lotus Japon...	shoot system	5	1016492091-1813157897-245964163-321305557-7
44	Mining for ...	44	32.00:00:00	2006-01-01 00:00:32	ia 75 mM N...	Lotus Japon...	shoot system	5	1622456557-552803112-2004561303-256861175-4
45	Mining for ...	45	32.00:00:00	2006-01-01 00:00:32	ia 75 mM N...	Lotus Japon...	shoot system	5	138989484-597336054-1607048547-1855428055-1

Figure 4: Experiment table on Mac.

3.2.4. Development of a Relational Data-Model

As science in general and metabolomics in particular experience a paradigm shift towards data driven research, the importance of metadata as an equivalent partner for actual observed metabolite profile levels is being increasingly recognized. Therefore, to tightly integrate XEML based experimental descriptions for metabolite profiles, we developed a relational data base schema for XEML data and extended the GMD accordingly.

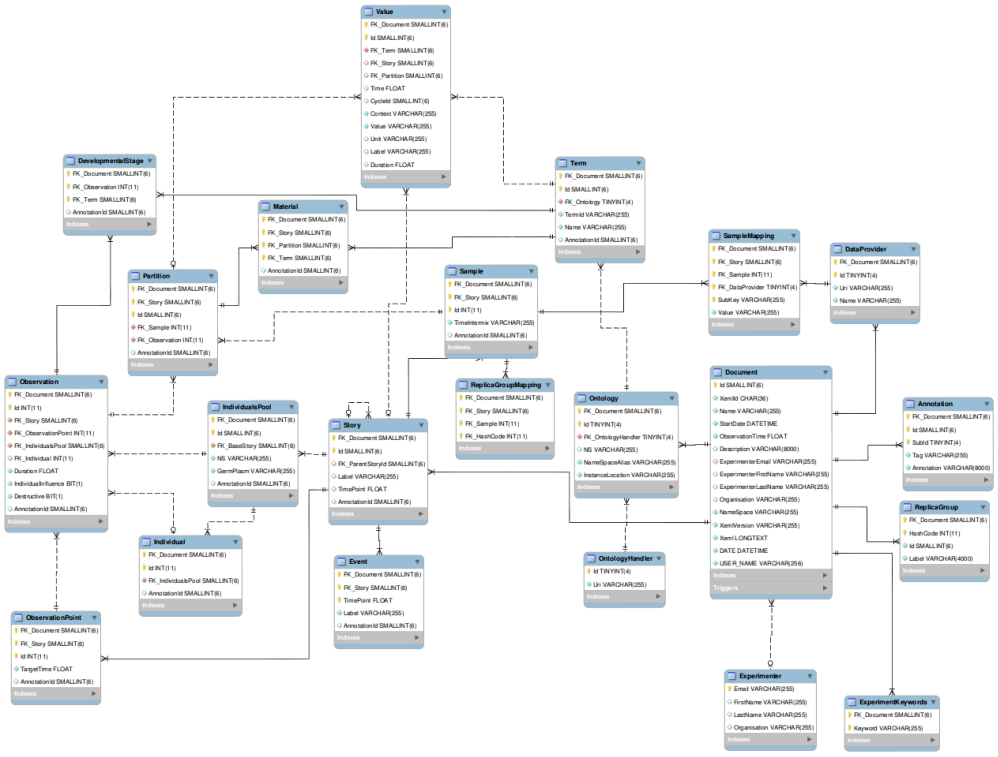


Figure 5: relational schema for XEML data

Decomposing the XEML based experimental description as shown in Figure 5 was also essential for the flexible creation of the ISA-Tab files. The schema provides the basic information for a successful mapping of data between the different concepts of the file formats. For example, the ISA-Tab format utilises Factors and Parameters to explicitly describe experimental conditions, with Factors kept constant and Parameters varying throughout an experiment. By contrast, XEML uses single Terms (a tuple of an ontology based identifier bound to an actual value) to implicitly annotate the experiment. The spectrum of values of single Terms is then used as a measure to decide whether a Term is understood as Parameter or Factor in ISA-Tab concept. Although both approaches are equivalent to each other, there is no simple language mapping tool (e.g. XSLT transformation) to convert one file format into the other. However,

as data integration and combined cross experiment data evaluation are crucial points on the COSMOS agenda, we evaluated detail of *implicit* vs. *explicit* annotation details further.

3.2.5. Implementation of a General Ontology Handler

Based on feedback from the community and the research conducted at the institutes involved in developing the XEML framework, we initially decided to focus on the structure capture of metadata in plant related experiments. To facilitated the adaptation to other research fields (e.g. medical, environmental, food sciences), we extended the XEML framework by implementing a general ontology handler, which enables the integration of any *obo*-formatted ontology into the XEML framework.

3.2.6. Implementation of a General Data Handler

To enable a mapping between samples specified/listed in XEML and actual experimental data (such as images, NMR spectrum or mass spectrometry chromatograms) on the user site, unique and specific identifiers must be assigned to both, samples and experimental data. As the samples are unambiguously enumerable within XEML, the identifiers for analytical data might depend upon the user's analytical pipelines and general settings (types of identifiers, number of sub keys in case of composed identifiers). To handle potentially all possible user environments, a general data provider handler was put into place. Neither programming skills nor knowledge of the underlying system are required from the user to map actual data files to XEML.

3.2.7. Ontology based Description of an GC-MS based Platforms

We collaborated with the COSMOS team working on deliverable 2.1 to extract a number of concepts and terms for the description of the analytical platform used in the GMD. They were collected in the COSMOS tracking system at Sourceforge (http://sourceforge.net/p/cosmos-fp7/tickets/), and discussed among the project partners. Finally, we compiled an example description for the *Varian5* GC-MS variant for initial exemplification and further discussion within the community.

protocol	property	value
extraction	water chloroform	(2:1; v/v) 15min at 70°C
derivatization	derivatisation	METHOXYAMINATION, 90min at 30°C
derivatization	derivatisation	TRIMETHYLSILYLATION, 30min at 37°C
separation	Ion source temperature	250°C
separation	Column length	30+10m (analytical column + guard column)
separation	Column inner diameter	0.25mm
separation	Temperature program	isotherm 1min 70°C, ramp 9°C/min, isotherm 5min 350°C
ion source	electron ionization	70eV
detector	LECO instrument model	Pegasus III TOF mass spectrometer
detector	highest observed m/z	m/z = 600
detector	lowest observed m/z	m/z = 70
detector	sampling frequency	20 scans/s
detector	mass analyzer type	MS-TOF
deconvolution	moving average smoothing	20
deconvolution	software	ChromaTOF
deconvolution	signal-to-noise ratio	2
RI	Retention index	decane
RI	Retention index	dodecane
RI	Retention index	pentadecane
RI	Retention index	octadecane
RI	Retention index	nonadecane
RI	Retention index	docosane
RI	Retention index	octacosane
RI	Retention index	dotriacontane
RI	Retention index	hexatriacontane
Chromatography	Chromatography	Agilent GC 6890

Figure 6: Screenshot taken from the GMD (http://gmd.mpimp-golm.mpg.de/Spectrums/1596e948-bff3-4eeb-b6ff-372085b06770.aspx) exemplifying best practice annotation of the actual GC-MS based analytical platform. The protocol column provides a general sub order of the terms. The property and value columns link to the actual ontology term pages (if available, indicated by an underlined link) where additional information might be found.

3.2.8. Export of Reference Data Sets to MetaboLights

Putting together the details reported in chapters 3.2.4 to 3.2.7, we were finally able to successfully orchestrate a semi-automated annotation pipeline. Metabolite profiling experiments can be designed, annotated, mapped to actual experimental data, and finally exported into ISA-Tab formatted data sets, which might be readily imported into the MetaboLights database. We used this pipeline to export metabolite-profiling experiments from the GMD into the MetaboLights database. Concerning the new version of XEML Designer, this step in still under development.

3.2.9. Development of Example Visualisations

To tackle the huge potential of different metabolomics approaches, we aimed to find minimal consensus in automated metabolite profile visualisations. Given the broad range of potential applications, informative metabolite profile visualisations need to be both, as generally as possible and as specific as necessary. We found network based visualisation and correlation heat maps to perform best for all-against-all visualisation. We put special attention into the automated and algorithmic generation of explicit sample/replica group descriptions from the implicit XEML meta-data.

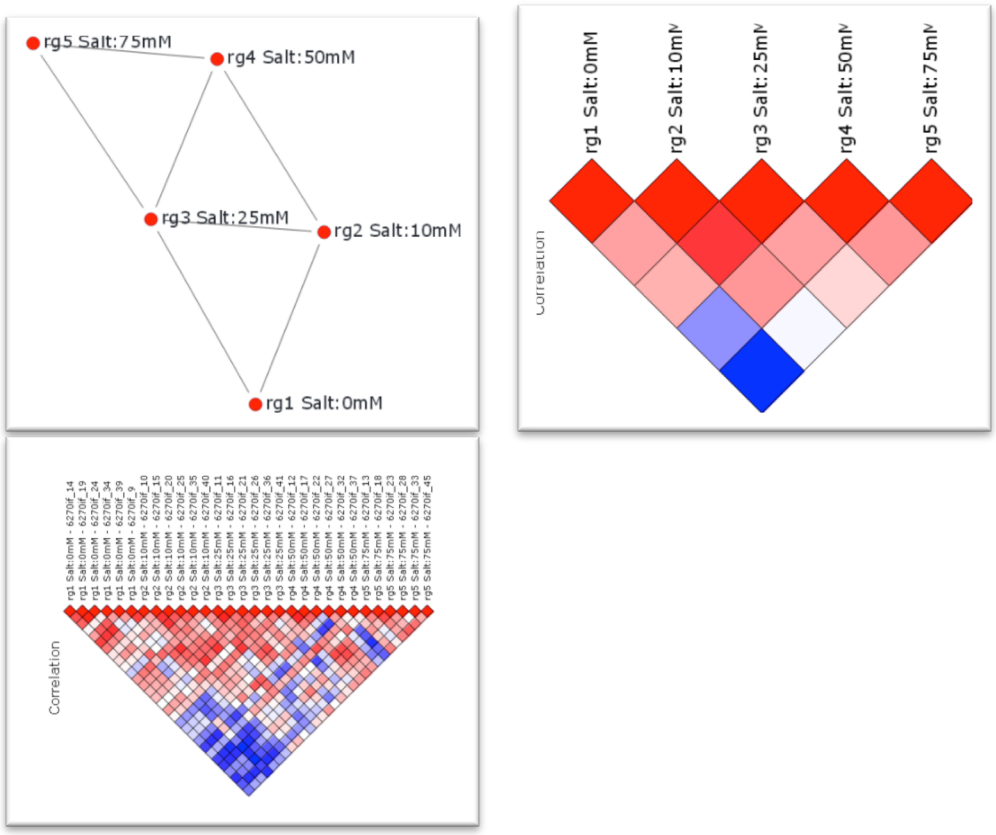


Figure 7: Screenshot of visualizations from correlations between experimental groups of an experiment. The plot top left encodes the strength of correlation within the distances between the nodes of a correlation network. The plot top right illustrates the strength of correlation colour-gradient-encoded (from red = high correlation over white = intermediate correlation to blue = no/weak correlation). From both visualisations can be seen that, in this example, the metabolite profiles of the experimental groups '*rg1 Salt:0mM*' are least correlated to group '*rg1 Salt:75mM*'. Equivalent visualisations might be used to represent correlations between individual samples (bottom left) or between the metabolites.

Name	Organism	Factor	Heatmap	Box plot	Variance	Anova F score
Mining for metabolic responses to long-term salt stress: a case study on the model legume Lotus japonicus (C)	Lotus	Salt			727.45	148.99
Metabolomic responses to long-term salt stress in related Lotus species (A)	Lotus	Germplasm, Salt			496.89	44.10
Metabolomic responses to long-term salt stress in related Lotus species (C)	Lotus	Germplasm, Salt			547.49	34.23
Mining for metabolic responses to long-term salt stress: a case study on the model legume Lotus japonicus (A)	Lotus	Salt			284.76	32.07
Metabolomic responses to long-term salt stress in related Lotus species (B)	Lotus	Germplasm, Salt			291.67	15.23

Figure 8: Heat map and bar plot visualisations of relative concentrations of a single metabolite across experimental groups of different legume experiments.

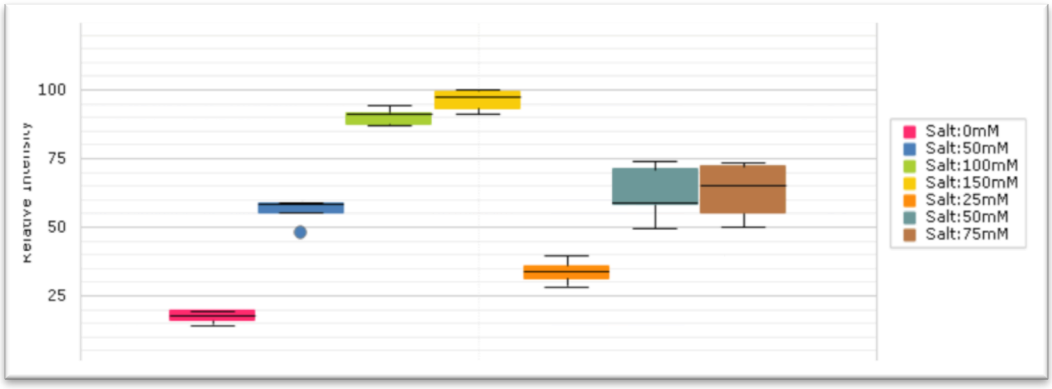
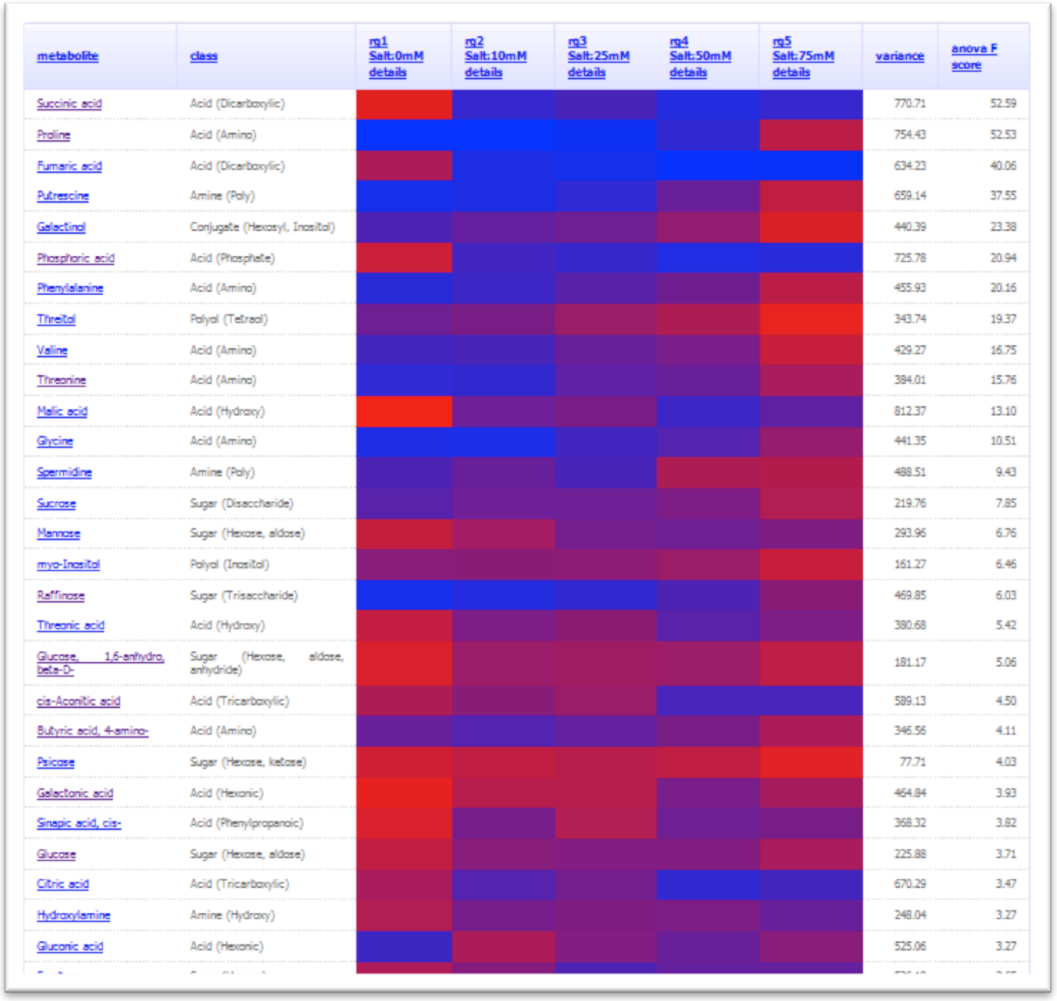


Figure 9: Box plot of relative metabolite concentrations in the replica groups of a single experiment.



metabolite	class	rg1 Salt:0mM details	rg2 Salt:10mM details	rg3 Salt:25mM details	rg4 Salt:50mM details	rg5 Salt:75mM details	variance	anova F score
Succinic acid	Acid (Dicarboxylic)						770.71	52.59
Proline	Acid (Amino)						754.43	52.53
Fumaric acid	Acid (Dicarboxylic)						634.23	40.06
Putrescine	Amine (Poly)						659.14	37.55
Galactinol	Conjugate (Hexosyl, Inositol)						440.39	23.38
Phosphoric acid	Acid (Phosphate)						725.78	20.94
Phenylalanine	Acid (Amino)						455.93	20.16
Threitol	Polyol (Tetraol)						343.74	19.37
Valine	Acid (Amino)						429.27	16.75
Threonine	Acid (Amino)						384.01	15.76
Malic acid	Acid (Hydroxy)						812.37	13.10
Glycine	Acid (Amino)						441.35	10.51
Spermidine	Amine (Poly)						488.51	9.43
Sucrose	Sugar (Disaccharide)						219.76	7.85
Mannose	Sugar (Hexose, aldose)						293.96	6.76
myo-Inositol	Polyol (Inositol)						161.27	6.46
Raffinose	Sugar (Trisaccharide)						469.85	6.03
Threonic acid	Acid (Hydroxy)						380.68	5.42
Glucose, 1,6-anhydro, beta-D-	Sugar (Hexose, aldose, anhydride)						181.17	5.06
cis-Aconitic acid	Acid (Tricarboxylic)						589.13	4.50
Butyric acid, 4-amino-	Acid (Amino)						346.56	4.11
Psicose	Sugar (Hexose, ketose)						77.71	4.03
Galactonic acid	Acid (Hexonic)						464.84	3.93
Sinapic acid, cis-	Acid (Phenylpropanoic)						368.32	3.82
Glucose	Sugar (Hexose, aldose)						225.88	3.71
Citric acid	Acid (Tricarboxylic)						670.29	3.47
Hydroxylamine	Amine (Hydroxy)						248.04	3.27
Gluconic acid	Acid (Hexonic)						525.06	3.27

Figure 10: Tabular visualization shows a profile of all identified metabolites from an individual *Arabidopsis thaliana* salt stress experiment. The table of metabolites is by default descending sorted according to an ANOVA F value and can alternatively be sorted by clicking the column headings.

3.2.10. Next steps

With this deliverable, we have developed a pipeline to process, store, and channel metabolomics measurement and associated experimental metadata into *MetaboLights*. We will continue to test, support and monitor the adoption of XEML and its subsequent use for pipelining data into *MetaboLights*. We will further extend the number of data provider (doing on-site visits) to integrate XEML with more stakeholder labs. Concerning the new XEML Designer and given that we have worked in close collaboration with the MetaboHub French National Infrastructure, we expect to produce a software widely used in Metabolomics and Physiology communities, and in particular the MetaboHub community. In a long-term perspective, we can imagine to extend the XEML Designer concepts and ontologies (first designed to describe plants experiments) to those that applicable to describe experiments for many other kinds of organisms. The connection to the PLATO pipepline and database and data imported/exported from "MetaboLights", coupled with the metadata captured in XEML Lab Designer, should also provide rich datasets for data mining and statistical analysis.

References

Hannemann, J., Poorter, H., Usadel, B., Bläsing, O.E., Fink, A., Tardieu, F., Atkin O.K., Pons, T., Stitt, M., Gibon, Y. (2009) Xeml Lab: A tool that supports the design of experiments at a graphical interface and generates computer-readable metadata files, which capture information about genotypes, growth conditions, environmental perturbations and sampling strategy. Plant Cell & Environment 32: 1185-1200

4. Publications

In progress.

5. Delivery and schedule

The delivery is delayed: ☐ Yes ☑ No

6. Adjustments made

No adjustments were made.

7. Efforts for this deliverable

Institute	Person-months (PM)		Period
	Actual	estimated	
8: MPIMP	10		Month 18
12: UB2	3		
1:EMBL-EBI	3		
2:LU	1		

7:UB	1		
3:MRC	0.5		
Total	18.5	25	

8. Appendices/ Links to software

1. Windows-XEML project websites: http://xeml.codeplex.com/, or http://gmd.mpimp-golm.mpg.de/apps/xemldesigner/
2. Ported XEML designer Sources and executable and tutorial (pdf) can be found at https://github.com/cbib/XEML-Lab

Appendices

3. 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	
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