Deliverable 5.1

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:	Tool that enables uploading of specific metadata to the MetabolomeXchange
WP No.	5
Lead Beneficiary:	2. LU
WP Title	Dissemination Pipelines
Contractual delivery date:	1 10 2014
Actual delivery date:	1 10 2014
WP leader:	Thomas Hankemeier 2. LU
Contributing partner(s):	1.EMBL-EBI, 8. MPI-MP, 11. IPB, 13. UB2, UCSD Metabolomics Workbench

Authors: Thomas Hankemeier, Christoph Steinbeck, Reza Salek, Kenneth Haug, Steffen Neumann, Theo Reijmers, Michael van Vliet

Contents

1 Executive summary 3
2 Project objectives 3
3 Detailed report on the deliverable 4
3.1 Background 4
3.2 Description of Work 4
3.2.1 Data set insert/update mechanism 4
3.2.2 Web interface 5
3.2.3 Access and documentation 6
3.3 Next steps 6
4 Publications 6
5 Delivery and schedule 6
6 Adjustments made 6
7 Efforts for this deliverable 7
Appendices 7
Background information 7

1 Executive summary

For this deliverable D5.1 we have coordinated the efforts from multiple international metabolomics data providers to make metabolomics data sets over their international data repositories searchable. We have designed and implemented a central online register called MetabolomeXchange to store meta-data of publicly available metabolomics data sets. With this central register we provide a search interface for finding data sets of interest that are available in the different data repositories. Data sets can be added or updated by the individual providers by updating their local data feed. The provider feed is then read by the MetabolomeXchange update mechanism and processed accordingly. We were able to connect all providers so far based on existing data feeds (XML/JSON) keeping technical and procedural changes to a minimum for the providers. To align the provider feeds we wrote feed converters to adapt the original feeds to MetabolomeXchange compatible feeds. In addition to the basic search we developed a 'popular searches' and 'recent searches' feature to improve the search experience.

2 Project objectives

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

No.	Objective	Yes	No
1	Provide a central online register of publicly available Metabolomics data sets called MetabolomeXchange.	X	
2	A mechanism to insert/update meta data of Metabolomics data sets by the individual data providers to MetabolomeXchange.	X	
3	A web interface to list and search for Metabolomics data sets available at MetabolomeXchange based on the provided meta data.	X	

3 Detailed report on the deliverable

3.1 Background

Over the last 5 years several metabolomics data repositories have been created. Numerous both special- and general-purpose repositories now exist making international collaborations and use and exchange of metabolomics data possible and easy. Examples of these repositories represented within the COSMOS consortium are:

- Metabolights, a general-purpose database for Metabolomics experiments and derived information that is cross-species, cross-technique and covers metabolite structures and their reference spectra as well as their biological roles, locations and concentrations, and experimental data from metabolic experiments.
- Metabolomics Workbench, a general-purpose, scalable and extensible informatics infrastructure that serves as the national metabolomics resource in the US, and which is funded by NIH.
- The Golm Metabolome Database (GMD) that facilitates the search for and dissemination of reference mass spectra from biologically active metabolites quantified using gas chromatography (GC) coupled to mass spectrometry (MS).
- MeRy-B, a plant metabolomics platform allowing the storage and visualisation of Nuclear Magnetic Resonance (NMR) metabolic profiles from plants.

3.2 Description of Work

Design and implementation of an infrastructure to support the data exchange of publicly available metabolomics data sets between data providers and the metabolomics community at large.

3.2.1 Data set insert/update mechanism

The original idea was that data providers would upload regularly updates to the MetabolomeXchange database. This however can be a tedious job, which requires a lot of manual steps and takes a fair bit of time to do, and may result in an incomplete capture of studies. After discussion with the persons who created

ProteomeXchange and talking to some of the bigger and more established metabolomics data providers we have chosen a different approach.

A pull mechanism seemed to be better fit for purpose as all providers we talked to already have some sort of data feed available with the information MetabolomeXchange required. We currently have four providers on board that we poll 4 times per hour to see if new or updated data sets are available by comparing feed checksums. If the checksum is different we parse the provider feed and process the changes. Because we decided to build on existing data feeds we had to convert the original provider feed to a MetabolomeXchange compatible feed. For each provider we now have a script that converts the original XML- or JSON-feed into a MetabolomeXchange compatible JSON-feed

3.2.2 Web interface

On top of the database we developed a web interface. It allows users to browse through and search for data sets of interest. To improve the search experience we added features like “popular searches” to see what others look for and a “recent searches” to keep track of your own recent searches.

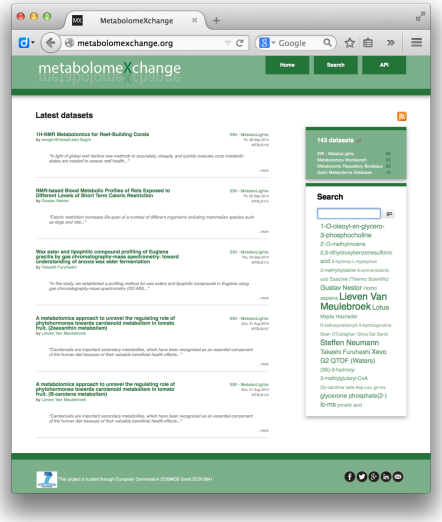

Figure A: Homepage showing latest data sets

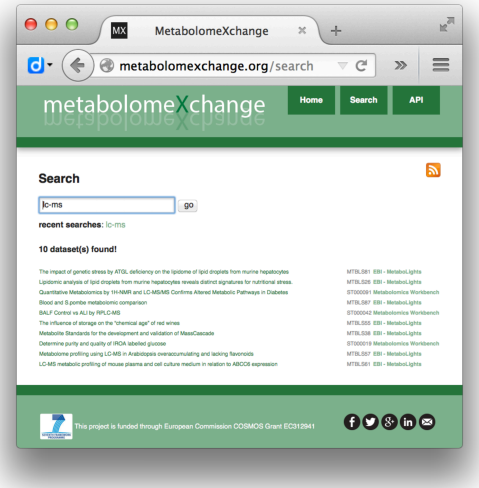

Figure B: Search page listing data sets of interest

3.2.3 Access and documentation

MetabolomeXchange is available and accessible at http://metabolomexchange.org. All source files are available on the project Github pages, together with accompanying readme files and license (Apache License, Version 2.0):

GitHub (application)**:** https://github.com/leidenuniv-lacdr-abs/metabolomexchange
GitHub (feeds)**:** https://github.com/leidenuniv-lacdr-abs/metabolomexchange-feeds

3.3 Next steps

Now that we have the first providers committed to share the metabolomics data set in their repository via MetabolomeExchange we will allow the community to define new features and improvements. We will now focus to create a stable platform and environment for data providers and the metabolomics community to exchange data. In order to achieve this we will try to align providers better in collaboration via WP4 “Data Deposition” and provide clear guidelines how providers share data within the context of MetabolomeXchange.

4 Publications

None.

5 Delivery and schedule

The delivery is delayed: ☐Yes ☑No

6 Adjustments made

- Name of system changed from MetaboStore to MetabolomeXchange.

- Instead of using an upload mechanism a feed aggregation (pull) mechanism has been build. This makes it possible to work without a provider specific login at application level making maintenance, now and in the future, easier and cheaper for both data providers and infrastructure maintainers.

7 Efforts for this deliverable

Institute	Person-months (PM)	
	actual	estimated
2: UL	7	
1: EMBL-EBI	1	
8:MPG	1	
11:IPB	0.5	
2:MRC	1	
UCSD Metabolomics Workbench	1 In kind	
Total	**10.5**	**12**

Appendices

1. N/A

Background information

This deliverable relates to WP5; background information on this WP as originally indicated in the description of work (DoW) is included below.

WP5 Title: Dissemination Pipelines
Lead: Thomas Hankemeier, UL
Participants: EBI-EMBL, LU-NMC, MRC, VTT, UB, MPG, IPB, UB2 and UBHam,

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	WP5	**Start date or starting event:**					Month 1		
Work package title	Dissemination Piplines								
Activity Type	COORD								
Participant number	1: EMBL-EBI	2: LU/NC	3:MRC	6:VTT	7:UB	8:MPG	11:IPB	12:UB2	13:UBHAM
Person-months per participant	7	15	2	2	3	2	1	1	2

Objectives

This work package will develop the mechanisms for disseminating the data submitted to all COSMOS partners to the other participating Metabolomics resources in the consortium, and the community at large. The desired setup will enable users to submit their data and metadata to any of the participating resources, whereupon it will be made available automatically to all other repositories or participants who wish to access the data, providing different, added value views of the data. Efficient user notification of new datasets and access to metadata will be provided through RSS notifications, and a central archive of such notifications. Reprocessed views of the data will also be announced and registered through this mechanism.

Description of work and role of participants

Task 1: Dissemination pipeline Once metabolomics data acquired by one of the COSMOS partners has been approved for public release (e.g. after assuring a certain quality level or after statistical analysis or publication), specific metadata will be automatically sent to all interested parties (all COSMOS partners and anyone interested in the metabolomics community) through RSS notifications. Checking the content of the metadata allows the receiver to decide if the dataset will be downloaded. The RSS feed does contain information (e.g. an URL) how to access the metabolomics data, possibly after checking authentication and authorization. The use-cases for this mechanism are manifold and of high interest to our user communities. One case would be experimentally derived standards. If a party is interested in a particular class of

compounds, say eicosanoids, it will be alerted whenever a new structure was submitted so an update of their local database can be triggered. Secondly, based on a grouping of metabolites according to tissue type, researchers interested in, for example, adipose tissue will be alerted whenever a new metabolite in adipose tissue is found. Finally this will have obvious benefits for any large-scale model organism studies - e.g. yeast, C. elegans, flies etc.
Task 2: Development of MetaboStore, a metadata archive for Metabolomics, serving as an intermediate general-purpose component to feed into the stakeholder repositories. In a later stage an RSS receiving party will be able to specify up front what kind of data is of their interest. A tool will be developed that will alert the interested party only after finding certain predefined information after processing the metadata. The same tool can be used to query over all COSMOS studies ever released to the public by searching the MetaboStore, a metadata archive for metabolomics data. Such a federated query could, e.g., together with semantic queries, relieve individual Databases from managing SOAP/REST/custom query interfaces. Standardized metadata together with WP3 allows querying over studies on sample level, metabolite level (identities), on quantitative level (content of the dataset, reference data), on statistical data analysis result level or certain combinations of these levels. TNO will give input on the development of biological relevant queries and will develop essential ontologies, to facilitate data exchange. With the standards defined in WP 2 and 4 this will actually be a phenotype database on metabolism, and will be embedded in large e-infrastructures such as ELIXIR and BioMedBridges to allow the data integration and interoperability with important European initiatives. The data warehouse within the LU/NMC-DSP, developed together with NuGO, consists of the generic study capturing framework (GSCF), a simple assay module (for clinical chemistry data) and a metabolite centric module, and is a candidate repository to store the relevant study (meta) data.
The user acceptance will be monitored through usage and download statistics provided by the source code management site of our choice (SourceForge/Google Code).
In addition we will perform surveys as part of the last two annual stakeholder meetings.

Deliverables

No.	Name	Due month
D 5.1	Tool that enables uploading of specific metadata to the MetaboStore	24
D5.2	Implemented data-broadcast mechanism	24
D5.3	Tool that allows checking predefined information in broadcast	30
D5.4	Tool that allows querying MetaboStore	30
D5.5	Usage statistic and downloads report	36