Deliverable D4.1

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:	COSMOS repository data flow definition: COSMOS repository data flow definition, as formally agreed by the members of the COSMOS consortium	
WP No.	WP4	
Lead Beneficiary:	THE UNIVERSITY OF MANCHESTER	
WP Title	Data Deposition	
Contractual delivery date:	01 July 2013	
Actual delivery date:	01 January 2014	
WP leader:	Roy Goodacre	UNIMAN

Contributing partner(s):	Elon Correa, Jan Hummel, Theo Reijmers, Philippe Rocca-Sera, Jules Griffin, Tim Ebbels, Marta Cascante ,Reza Salek, Roy Goodacre

***Authors**: Elon Correa, Jan Hummel, Theo Reijmers*, *Reza Salek and Roy Goodacre*.

Contents

1 Executive summary 3
2 Project objectives 3
3 Detailed report on the deliverable 3
3.1 Background 3
3.2 Description of Work 4
3.3 Next steps 5
4 Publications 6
5 Delivery and schedule 6
6 Adjustments made 6
7 Efforts for this deliverable 6
Appendices 7

1 Executive summary

The aim of this deliverable is to propose a guideline for data deposition workflow between potential and participating metabolomics databases and repositories. This would ensure a coherent metabolomics workflow to runs to its full potential, capturing agreed sets of metadata across different resources. The workflow definitions will prioritise simplicity, usability, annotation quality and the plurality of metabolomics resources and databases to ensure a coherent connectivity between similar studies and to provide rapid matching results to end users.

2 Project objectives

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

No.	Objective	Yes	No
1	Definition and implementation of deposition data flow in the COSMOS consortium	X	
2	Define the joint COSMOS data format and submission requirements		X

3 Detailed report on the deliverable

3.1 Background

Due to the complexity of chemical processes involving metabolites and the high-throughput, diversity and sensitivity of various analytical methods used in metabolomics, this field generates vast amounts of raw data and require subsequent biological and statistical analysis to understand the results. Making raw data, post-processing methods, statistical methods and source codes available to the interested research community has clear benefits to the transparency and trustiness of the scientific studies results promoting further data peer-reviewing, replication and validation of the findings. The COSMOS data flow guidelines will ensure a cross resource access to various resources, protecting data proprietary interests, security and confidentiality as required.

3.2 Description of Work

COSMOS will establish clear procedures for metabolomics data submission and deposition, results reporting and publishing requirements. This will ensure proper reporting of metabolomics data, metadata, annotation and that required minimum information is captured according to the existing Metabolomics Standards Initiative (MSI) guidelines. These new guidelines are currently being carefully discussed, elaborated and agreed by all COSMOS partners. COSMOS is also taking every opportunity to engage with stakeholders and potential collaborators on planning, discussion and implementation of the guidelines for data deposition workflows. Several of the COSMOS consortium participants are Members and Directors of the Metabolomics Society, also on the Board of other "omics" standardization initiatives, ensuring links and cross talks, and working with publishers. For example, new data publication platforms such Nature Publishing Group's Scientific Data and BioMedCentral/BGI's GigaScience already use the ISA framework adapted as a means of capture metabolomics metadata in MetaboLights.

In September 2012, National Institutes of Health (NIH) Common Funds Metabolomics program awarded funding related to metabolomics research advancement, funding three Regional Comprehensive Metabolomics Research Cores (RCMRC) and a Data Repository and Coordination Centre (DRCC) to act as a North American hub for metabolomics related research [1]. A second round of proposals is currently under evaluation. During the COSMOS stakeholder meeting in Glasgow (July, 2013) one of the main outcomes was to plan a joint meeting at EMBL-EBI in the 4th quarter of 2013. This meeting will mainly be between MetaboLights, the EMBL-EBI general-purpose open source metabolomics repository, and the NIH metabolomics initiatives and aim to work towards a set of agreeable metadata workflow exchanges and ways to share data and resources.

At the time of writing, no precise workflow has been established but a proposed model for the data deposition workflow (Figure 1) has been drafted within COSMOS. The data deposition cycle is initiated when a submitter (who has generated or owns the study material) submits their metabolomics study to a specific associated database (e.g. Metabolights, Netherlands Metabolomics Centre database, Golm Metabolomics Database, …). Once the data submission has fulfilled the metadata-reporting requirement of the associated repository, a unique COSMOS accession number will be generated.

The "*COSMOS engine/website*" similar to the proteomics (proteomexchange.org) will then properly annotate, format and store the minimum agreed metadata according to the proposed reporting standards suggested by work packages 1, 2, 3 & 5. This proposal is currently under discussion with collaboration partners, metabolomics repositories and stakeholders. Once all data and information acquired has been deposited into an associated metabolomics database, such as MetaboLights, an automatic reporting would be generated based on agreed minimum metadata information (D4.2), along with the unique accession number. This information would then be displayed via a proposed build web application. We envisage that in the future, additional purposely-built databases can potentially be integrated into this proposed workflow.

The first phase of the data deposition cycle is temporary and all data and associated information are kept private. However, if the study has been submitted for publication, the depositor may authorize reviewers (or journal) to access the data via unique COSMOS accession number data link (a temporary link) to the where the data has been deposited. This is mutually agreed between the respective COSMOS partner and the publishing journal involved. Once the depositor agrees to make the data open access and the study has been officially published, the COSMOS system will automatically make the study freely available to the broader research community. All parties involved will greatly benefit from sharing raw data, metadata, statistical methods and source code, thereby ensuring that the whole scientific process is more transparent. By increasing the visibility of their work, depositors are likely to boost citations. The publishing companies and journals will expose their publications to a greater number of potential readers and enhance impact factor. In addition, through COSMOS the research community will gain free access to a vast amount of well documented scientific information.

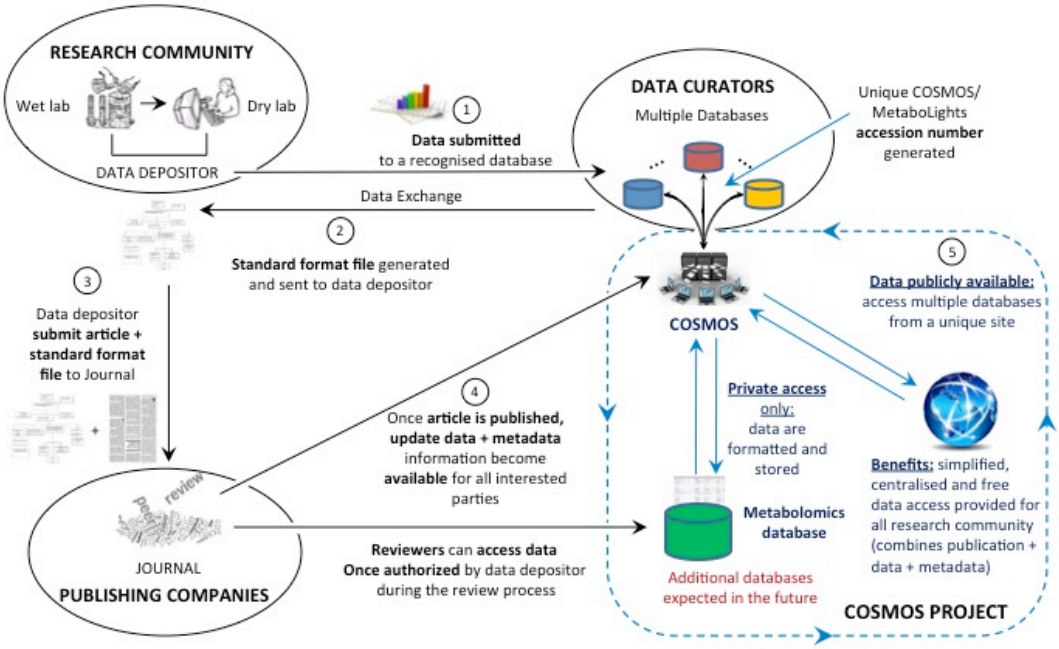

Figure 1: Initial draft model for the COSMOS data deposition workflow.

3.3 Next steps

COSMOS will bring together publishers, journals and metabolomics repositories such as the NIH data centres (Metabolomics Workbench), Netherlands Metabolomics Centre (NMC) [2] and the Golm Metabolome Database (GMD) [3], amongst others for a final agreement on data workflows, minimum metadata reporting on associated raw data, source code and any additional information that

will benefit the community through a shared model. Open access to the system will ensure that any interested party can benefit from its full capabilities.

The final decisions regarding specific data repositories accessed and the role and involvement of publishing companies in the data deposition flow are still being discussed with partners, stakeholders and collaborators. Our objective is to gather as sufficient information before making any final decisions. This will involve further meetings and discussions. Careful planning of the data deposition workflow, its control policies and actions will ensure that the utilisation and adaptation of the system is maximised for use inside and outside Europe.

4 Publications

N/A

5 Delivery and schedule

The delivery is delayed: ☑ Yes ☐ No

As the objective is to maximise the applicability of the system worldwide, existing and potential collaborators are being engaged in the decision process. This has unfortunately delayed the formal definition as further discussions are required to process the input of all interested parties. This diverse engagement will ensure that the results of COSMOS are of maximal use for the community and beyond.

6 Adjustments made

We plan to extend this deliverable by 6 months to give sufficient time for an agreeable and comprehensive data workflow between existing COSMOS partners and stakeholders and the NIH.

7 Efforts for this deliverable

Institute	Person-months (PM)		Period

	actual	estimated	
9: UNIMAN	2		9
2:LU/NMC	1 (in Kind)		
1:EMBL-EBI	0.5		
7:UB	1		
8:MPG	0.5		
14:UOXF	0.5 (in Kind)		
2:MRC	1.39		
4:IMPERIAL	0.12		
Total	7.01	9	

Appendices

1- NIH News:*NIH announces new program in metabolomics*. 2012. http://www.nih.gov/news/health/sep2012/od-19.htm.

2- Netherlands Metabolomics Centre; http://www.metabolomicscentre.nl/

3- The Golm Metabolome Database (GMD); http://gmd.mpimp-golm.mpg.de/

Background information

This deliverable relates to WP4; background information on this WP as originally indicated in the description of work (DoW) is included below.

WP4 Title: Data Deposition
Lead: Roy Goodacre, UNIVERSITY OF MANCHESTER
Participants: WP1, WP2, WP3 and WP5

First, we will implement harmonized and compatible data deposition and annotation strategies across all partners, providing data producers involved in Metabolomics experiments with a single point of submission. The data deposition and exchange workflow in the COSMOS consortium will be formally defined, agreed, and documented in relation with WP3 and all partnering databases in Europe and world-wide that will be invited to participate.
As a second objective, we will work towards the generation of an annotation manual for submitted data and strive to make sure that all metabolomics data submitted to partner databases are annotated to this standard.
Since the adoption of minimal standards for metabolomics by the relevant journals is a major goal of this coordination action, we are going to consult with publication houses and ensure data annotation quality and consistency, according to the required standard level set by each journal.
In this activity the work by the BioSharing initiative (http://biosharing.org) will also be explored. Building on the effort of Minimum Information for Biological and Biomedical Investigations' (MIBBI) portal (http://mibbi.org), the BioSharing initiative works to strengthen collaborations between researchers, funders, industry and journals, and to discourage redundant (if unintentional) competition between standards-generating groups.

Work package number	**WP4**	**Start date or starting event:**									month 1
Work package title	Data Deposition										
Activity Type	Coord										
Participant number	1: EMBL-EBI	2: LU/NMC	3: MRC	4:mperial	6: VTT	7: UB	8:MPG	9: UNIMAN	11:IPB	12: UB2	13:UBHAM
Person-months per participant	9	6	6	6	2	2	2	14	1	2	2

Objectives

1. First, we will implement harmonized and compatible data deposition and annotation strategies across all partners, providing data producers involved in

Metabolomics experiments with a single point of submission. The data deposition and exchange workflow in the COSMOS consortium will be formally defined, agreed, and documented in relation with WP3 and all partnering databases in Europe and world-wide that will be invited to participate.

2. As a second objective, we will work towards the generation of an annotation manual for submitted data and strive to make sure that all metabolomics data submitted to partner databases are annotated to this standard. Since the adoption of minimal standards for metabolomics by the relevant journals is a major goal of this coordination action, we are going to consult with publication houses and ensure data annotation quality and consistency, according to the required standard level set by each journal.
3. In this activity the work by the BioSharing initiative (http://biosharing.org) will also be explored. Building on the effort of Minimum Information for Biological and Biomedical Investigations' (MIBBI) portal (http://mibbi.org), the BioSharing initiative works to strengthen collaborations between researchers, funders, industry and journals, and to discourage redundant (if unintentional) competition between standards-generating groups.

Description of work and role of participants

Task 1: Definition and implementation of deposition data flow in the COSMOS consortium. The value of metabolomics data without proper biological, technical and statistical background is really quite limited. This was recognized by the Metabolomics Standards Initiative (MSI) and this resulted in a series of guidelines for minimum reporting standards that should be used for metabolomics experimentation (published in *Metabolomics* **3(3)** in 2007). In a close collaboration of all COSMOS participants, and after consultation with stakeholders (viz. MSI, Metabolomics Society, relevant Publishers, National and international funders), we will define the COSMOS data deposition workflow. MSI guidelines will be followed and we shall co-ordinate the representation of results and metadata in a relational database/XML representation, with data stored as WP2-compliant formats. We will define the joint COSMOS data format and submission requirements, likely a thin metadata wrapper around MSI data formats. On successful submission, a standard format file will be generated, containing a COSMOS accession number, metadata, and a private data access option for the use of the data owner and reviewers. The file will be sent to the data depositor, for him/her to pass on to the journal for review purposes. On publication of a manuscript, the associated dataset will be released by publisher and/or corresponding author, and an updated version of the metadata will be issued via the COSMOS RSS notification system, allowing all interested parties to access, process, and import the relevant data. This will have tremendous benefit to the metabolomics community, allowing others to re-create statistical approaches, providing data for others to mine and allowing the peer review process to access the raw and processed data of an experiment. The precise format of this has not yet been implemented and as discussed above we shall engage all stakeholders as well as publication houses. This task involves contributions from all COSMOS participants to deposit data and test the validity of the developed workflows, reflecting the central role of the data deposition workflow for all partners involved.

Task 2: Implementation of a MSI journal validation system. As discussed in Task 1 the value of metabolomics data without proper biological, technical and statistical background is really quite limited. This task will develop tools to validate compliance of the submitted metabolomics data with the MSI guidelines or specific journal requirements. This is not meant to tell people how to perform their analyses but to allow adequate reporting of what was performed so that others can repeat the work. As a result of the validation process, after COSMOS data deposition, a report about guideline compliancy of each submission will be generated automatically. This would aid Reviewers of articles submitted for publication as well as Editors handling paper submissions. Springer will pilot this initial system as the publisher of *Metabolomics* (http://www.springer.com/life+sciences/biochemistry+%26+biophysics/journal/11306) with the backing of the International Metabolomics Society (http://www.metabolomicssociety.org/) as this is their official journal. Several of the COSMOS consortium participants are Members and Directors of the Metabolomics Society. In addition many other journals are interested in developments in this area including *Nature Biotechnology* (Nature PG), *Genome Biology* (BMC), *Molecular Systems Biology* (RSC) and *Molecular BioSystems* (Nature PG and EMBO).

Deliverables

No.	Name	Due month
D4.1	COSMOS repository data flow definition	9
D4.2	COSMOS metadata format definition	9
D4.3	MSI implementation of the COSMOS data flow	15
D4.4	Consultation of the MSI implementation of the COSMOS data flow Publishers and International Society	15
D4.5	Implementation of MSI/journal validation system	15