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:	1 April 2014 (postponed by 9 months as agreed by all partners)	
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, Michael van Vliet, Reza Salek and Roy Goodacre*.

Contents

1 Executive summary

The aim of this deliverable is to define guidelines for data deposition workflow between participating and potential metabolomics databases and repositories. This will ensure a coherent metabolomics workflow to run 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. In collaboration with stakeholders, member of metabolomics society, publishers and partners, appropriate strategies for the sustainability of the data deposition workflow are also being discussed.

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. The general data flow commonly agreed by stakeholders, publishers and COSMOS' partners is depicted in Figure 1. The data flow is described in 4 stages, 3 of which directly communicate with the COSMOS data flow control system. Each of these stages is described below.

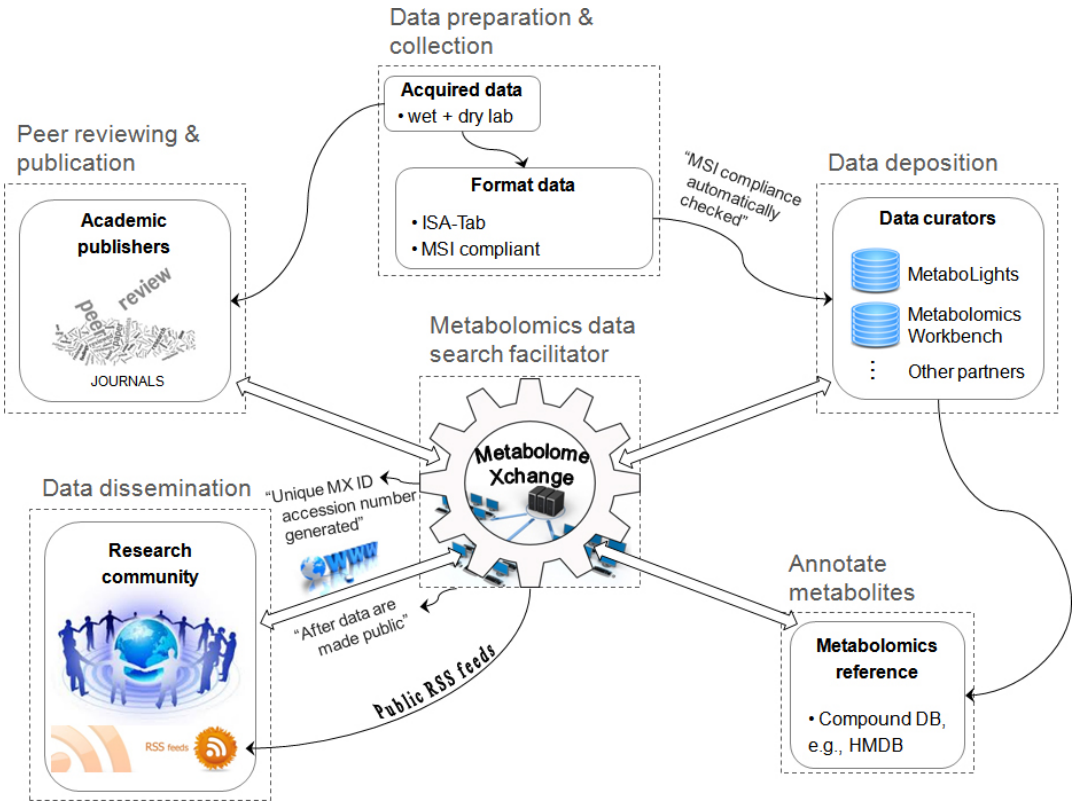
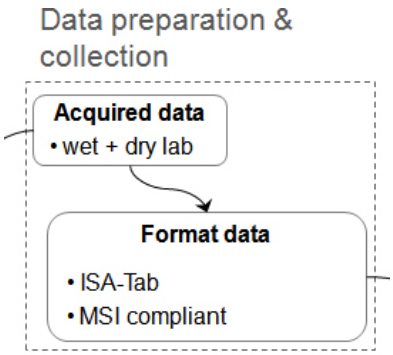


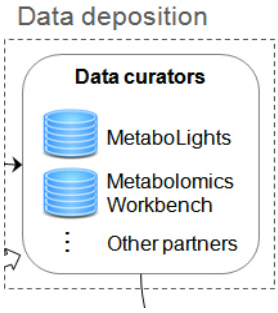
Figure 1: Current COSMOS data deposition workflow model.

3.2.1 Data preparation & collection



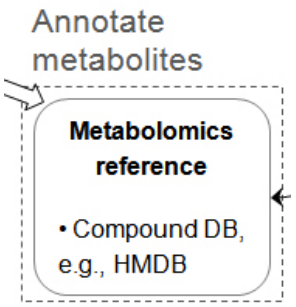
This stage refers to the data preparation and collection starting with the basics: data generation. The data acquisition, prior to the start of the COSMOS data flow, is based on a typical metabolomics data generation scenario where, given a hypothesis or a research problem, samples are collected and experimental data (e.g., GC-MS, spectroscopy, etc.) are generated (wet lab). The data are then usually preprocessed and statistically analysed (dry lab). The data depositor then submits experimental data, plus metadata, to an open data repository using a metadata annotation tool, such as ISA-creator (ISA-Tab), to be formatted according to community agreed standards following the MSI guidelines.

3.2.2 Data deposition



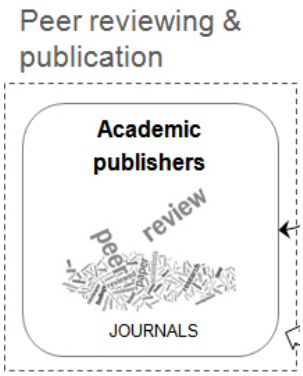
Once the data are MSI compliant the data producer (e.g. researcher) submits the data to one of the appropriate partner metabolomics database where the data and metadata will be stored together. However note, that the standards reporting requirement would be dependent on local policy of each repertory. Once the data has been processed, checked and approved for submission, a report will be generated on completion containing the minimum metadata identifying the study (Details in WP4, D4.2). This information from the respective metabolomics repository will then be pushed to MetabolomeXchange and subsequently becomes publically available once the submitter controlled embargo date is reached. At the time of data submission, MetabolomeXchange will automatically assign a unique accession number (ID) identifying the data set for further reference in the MX system.

3.2.3 Data annotation



Once the data are submitted, further automated or manual curation of data will annotate the reported metabolite using well-known and established databases such as ChEBI, LipidMaps or the Human Metabolome Database (HMDB). Such reference data resources will also be linked to MetabolomeXchange for community awareness and announcement of new data set availability.

3.2.4 Peer reviewing & publication



After confirmation that the depositor meet all data submission requirements, a related article can be submitted to one of the partner journals. One of the benefits for the partner journal is that the data and metadata have already been checked for compliance and meet the community agreed data requirements. With optional previous authorisation (e.g. Reviewer access or account) given by the data depositor, the data may be made available to the respective journal reviewers for inspection and clarification if needed. However, data proprietary interests, security and confidentiality will always be respected.

3.2.5 Data dissemination

After a submitted study has been accepted and made public the associated data and metadata will be made publically available to the research community via the MetabolomeXchange system. For instance, suppose that a submitted study describes disease *Y1* and has been submitted as an article entitled *X1* associated to data *D1* (deposited in a partner database). If the researcher (end user) enters a search for disease "*Y1*" into the MetabolomeXchange search engine, one of the results returned would be: article *X1* associated to data *D1*, including links to retrieve both data (freely) and article (freely or not depending on publisher policies). MetabolomeXchange will also periodically send RSS feeds to subscribed users informing of new relevant data and studies available.

3.3 Next steps

The effective dissemination and exploitation of the MetabolomeXchange-COSMOS project and the policies developed will be a key to a sustainable implementation of COSMOS. The next subsections refer to the latest Meeting on International Data Exchange in Metabolomics co-organized and sponsored by the international Metabolomics Society and COSMOS initiative which took place in the EMBL-EBI in Cambridge UK on the 2nd of April 2014. This meeting brought together stakeholders, publishers and COSMOS partners including representatives from the MSI, the Metabolomics Society, the *Metabolomics* journal, *Nature Publishing Group* and databases from Europe, the USA, Canada, Asia and Australia. The list of participants was as follows:

Present: Merlijn van Rijswijk (NMC), David Wishart (Univ. Alberta), Dirk Walter (MPMPI), Joachim Kopka (MPiMPI), Lloyd Sumner (Noble Fnd), Susanna Sansone (Un. of Oxford, Nature Publishing Group), Leslie Derr (NIH), Christoph Steinbeck (EBI), Masanori Arita (Nat. Inst. Genetics/Mass Bank), Phil Smith (NIH), Shankar Subramanian (UCSD), Oliver Fiehn (UC Davis), Rick Dunn (Univ. Birmingham), Mark Viant (Univ. Birmingham), Roy Goodacre, (Univ. Manchester), Art Castle (NIH), Ken Haug (EBI), Reza Salek (EBI), James Smith (MRC-HNR) and via Skype video call Saravanan Metabolomics Australia.

The outcome and agreements reached during this meeting are documented in the minutes of the meeting, which will become publically available via COSMOS website.

3.3.1 Sustainability: data sharing post COSMOS project

During the above-mentioned meeting the following agreement has been reached regarding data sharing sustainability:

- To provide a network of stable, coordinated, freely accessible metabolomics data from repositories that handle public submissions.
- To jointly make all published metabolomics data easily accessible for the scientific as well as commercial user community.
- To work closely with publishers, instrument vendors, software developers, data generation facilities, MSI, Metabolomics and the user community in the field of metabolomics to promote data accessibility.
- Decisions about procedures and membership will be made by the database providers in the MetabolomeXchange, after consulting all members of the consortium.
- Database Providers must implement "public" and "private" data access mechanisms.
- Once released, all data must be and remain fully freely and publicly accessible to all potential user groups, without additional steps like user registration or limitation of access (There may be some exceptions related to clinical data with personal identifiers).
- Database Providers may leave the Consortium at any time by notification to the other Database Providers.
- Leaving Database Providers must make all their data records available for import by an active partner database, for a 12 month period following departure, such that they may continue to be made searchable via the MetabolomeXchange portal (changing the underlying URLs). The importing database will then actively maintain these records but will acknowledge the originating database within the record.

In addition, in COSMOS data sustainability will be based on the EBI data sustainability policies and the ELIXIR project (http://www.elixir-europe.org/) whose objective is to provide the facilities necessary to store and share data for life by building a sustainable European infrastructure for biological information, supporting life science research and its translation to medicine, agriculture, bioindustries and society.

3.3.2 Feedback from stakeholders, publishers and final users

As it has been standard in COSMOS (and is evidentiated in section 3.1), the opinions of stakeholders, publishers, partners and final users are being heard and documented via joint collaborative meetings, stakeholders meeting, workshops, conference and specialized meetings. The results and meeting minute has been

reported via WP7, deliverables and made publically available via COSMOS website (http://cosmos-fp7.eu/wp7)

3.3.3 The use of standard data formats as developed in WP2

COSMOS actively participated in the 2014 Proteomics Standards Initiative (PSI) meeting (13-16th April, 2014) held near Frankfurt, Germany, where community standards for data representation in proteomics to facilitate data comparisons, exchange and verification have been extensively discussed. This meeting was sponsored by the BBSRC BBR grant “PROCESS” (code BB/K01997X/1), by the EU FP7 grants “ProteomeXchange” (no. 260558) and “COSMOS” project (no. 312941). During this meeting, the best practices and standard data formats for proteomics data sharing have been debated. For Metabolomics data, COSMOS will adopt the formats defined by PSI and descried and implemented by WP2. The presentation and the outcome of the meeting is available via HUPO-PSI website (http://www.psidev.info/ and http://www.psidev.info/psi2014)

3.3.4 Measuring the success of the work involved in WP4

Although WP4 is tightly connected to other work packages, in particular with WP2 & WP3 and WP5 at a large degree its success can be recognized independently from other WPs. Arguably, WP4’s biggest achievement so far has been to bring all interested parties together (stakeholders, publishers, partners and end users) and in common agreement elaborate a detailed and complex data flow for the project (see, Figure 1). In term of collaboration, organization and plurality of opinions considered, WP4 already is a success story. On the other hand, a more practical measure of success, which will evaluate the whole project, will only be possible when the COSMOS MetabolomeXchange system is fully implemented and is evaluated by the end users.

4 Publications

- Salek, R.M., Steinbeck, C., Goodacre, R., Viant, M.R. & Dunn, W.B. (2013) The role of reporting standards for metabolite annotation and identification in metabolomic studies. *GigaScience* **2**: 13
- Kirsten Gracie, Elon Correa, Samuel Mabbott, Jennifer A. Dougan, Duncan Graham, Royston Goodacre and Karen Faulds. “Simultaneous detection and quantification of three bacterial meningitis pathogens by SERS”. *Chem. Sci.*, 2014, 5, 1030.

5 Delivery and schedule

The delivery is delayed: ☑ Yes ☐ No

6 Adjustments made

None

7 Efforts for this deliverable

Institute	Person-months (PM)		Period
	actual	estimated	
9: UNIMAN	5	9	9
2:LU/NMC	1		
1:EMBL-EBI	2		
8:MPG	1		
Total	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