

## Deliverable D4.3

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:	MSI implementation of the COSMOS data flow	
WP No.	WP4	
Lead Beneficiary:	THE UNIVERSITY OF MANCHESTER	
WP Title	Data Deposition	
Contractual delivery date:	31 December 2013	
Actual delivery date:	31 December 2013	
WP leader:	Roy Goodacre	UNIMAN
Contributing partner(s):	Partner name(s) Elon Correa, Dirk Walter, (WP3) Thomas Hankemeier (WP5), Steffen Neumann, Daniel Schober (WP2), Reza Salek (WP1), Roy Goodacre	



*Authors: Elon Correa, Reza Salek, 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.....	11
4	Publications .....	11
5	Delivery and schedule .....	11
6	Adjustments made .....	11
7	Efforts for this deliverable.....	12
8	References .....	12
	Background information.....	14



# 1 Executive summary

This deliverable aims to describe actions taken towards the effective implementation of Metabolomics Standards Initiative (MSI) (Fiehn et al. 2007) standards to the COSMOS data flow. To achieve this common schema, a number of community agreed standards, or attributes, are being extracted from MSI and carefully discussed and evaluated by all COSMOS members. The objective is to deliver a COSMOS data flow which communicates data efficiently and effectively and is as MSI compliant as possible.

## 2 Project objectives

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

No.	Objective	Yes	No
1	MSI implementation of the COSMOS data flow	X	

## 3 Detailed report on the deliverable

### 3.1 Background

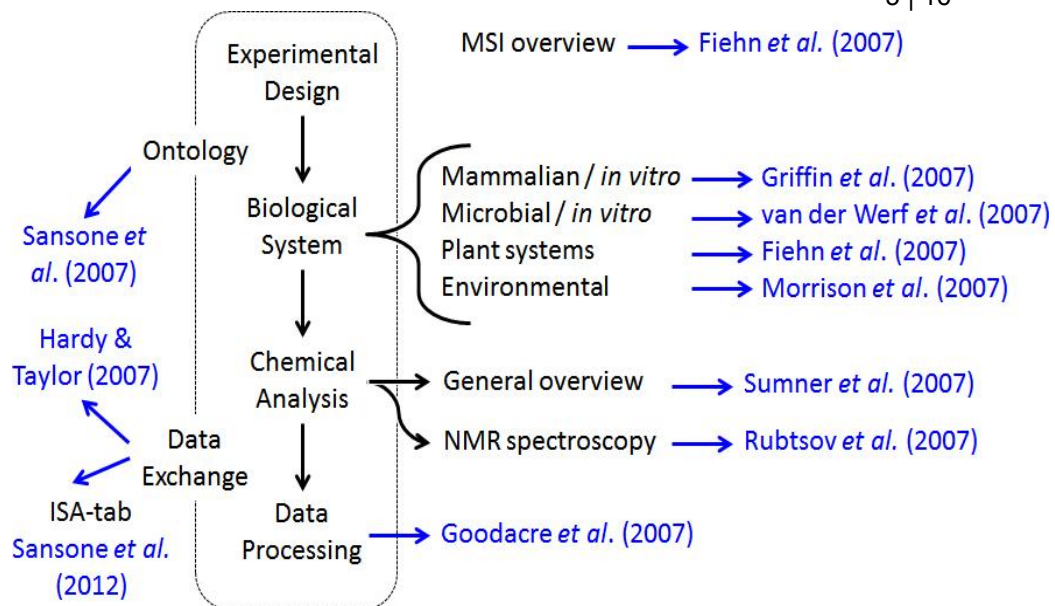
Well documented datasets freely available and shared with the scientific community for secondary data usage will add value to existing knowledge and accelerate advances in metabolomics. To drive these changes, the development and implementation of an efficient e-infrastructure is a necessity. Such platform will harmonize the data-flow using clearly defined communication protocols, data exchange standards and minimum information (MI) reporting. In 2005 the



Metabolomics Standards Initiative (MSI) (Fiehn *et al.* 2007, Sansone *et al.* 2007) was conceived following earlier work by the Standard Metabolic Reporting Structure (SMRS) initiative and the Architecture for Metabolomics (ArMet) consortium. The early efforts of MSI were focused on community-agreed reporting standards, which defined the minimum information required for the clear description of the biological system studied and all components of metabolomics studies. These MI standards are currently collected under the biosharing umbrella (<http://www.biosharing.org/standards/mibbi>). Collectively, the MSI workgroups (WG) are working towards developing the following *Core Information for Metabolomics Reporting*. This document will specify the minimal guidelines reporting metabolomics work. It will do so in a textual form and will seek in the long term to cover all application areas and analysis technologies. This document will be developed by the biological context metadata working group (WG), the chemical analysis WG, the data processing WG, the exchange format WG and the ontology WG.

### 3.2 Description of Work

The true relevance of MSI for data sharing and experimental protocols in metabolomics was recently summarized in an editorial (Goodacre 2014) which listed all community-agreed MI papers relevant in the metabolomics domain. Figure 1, extracted from the editorial, connects each of those MI standards to the topic domain at which they are aimed.



**Figure 1:** Overview of the metabolomics data pipeline showing where existing MSI minimal reporting guidelines (highlighted in blue) are aimed.

Some of these MI standards have also been solicited by major journals to enforce policies as part of publishing *best practices*. The Sumner *et al.* (2007) reporting standard on metabolite identification levels for example was made a prerequisite to submit to the *Metabolomics* journal, which now requires its authors to follow these guidelines explicitly, and tables with lists of metabolites must now contain reference to which of the four community agreed levels of metabolite identification have been made (these are provided in Table 1).



**Table 1:** A summary of the four-metabolite identification levels as defined by MSI (Sumner et al. 2007).

1	Identified compounds
	<p>This definitive identification level requires a minimum of at least two independent and orthogonal data relative to an authentic compound analysed under identical experimental conditions in the same laboratory on the same analytical platform</p> <p>These orthogonal data must provide different physicochemical properties of the metabolite and for examples may be: retention time/ index and mass spectrum; retention time and <math>^1\text{H}/^{13}\text{C}</math> NMR spectrum</p>
2	Putatively annotated compounds
	<p>These are very similar to level 1 BUT are identifications that are made without chemical reference standards. The above orthogonal characteristics are still used but are typically matched against public or commercially available spectral libraries</p>
3	Putatively characterized compound classes
	<p>This level defines compounds that are based upon characteristic physicochemical properties of a chemical class of compounds, or by spectral similarity to known compounds of a chemical class</p> <p>For example, a fatty acid like C18:1 fatty acid where the unsaturation point is not known (for oleic acid–(9Z)-octadec-9-enoic acid– this would become C18:1 cis-9 if matched with a standard). Another example is a C6 sugar</p>
4	Unknown compounds
	<p>Whilst these are unidentified or unclassified compounds, these small molecules can still be differentiated, recognized again by the analyst in further analysis, and therefore quantified based upon spectral data</p>



COSMOS is establishing clear procedures for metabolomics data annotation, submission and deposition, results reporting and publishing requirements. The *Metabolomics* journal, for instance, is already implementing procedures to encourage authors to comply with MSI standards and make data available to reviewers, during the review process, and the general community (via MetaboLights, MSI ISAConfiguration setup for instance) after publication. By creating an **information highway** for standardized metabolomics data, the COSMOS WP 4 will ensure proper storage, distribution and re-use of MSI compliant metabolomics data. We will establish communication channels which allow the use of MSI compliant exchange syntaxes and which allow to be quality-assured by corresponding minimum information guidelines. These new guidelines are currently being carefully discussed, elaborated, adapted and agreed by COSMOS partners, data curators, metabolomics society and publishers.

The consortium is also taking every opportunity to engage with stakeholders and funding bodies, such as RCUK, EU, other national councils, US NIH and potential collaborators on planning, discussion and implementation of the MSI guidelines for data flow. Several of the COSMOS partners are members and directors of the Metabolomics Society, and also on the Board of other “omics” standardization initiatives, ensuring links and cross talks. COSMOS is also in close contact with publishers. For the spreadsheet-based annotation of experimental metadata COSMOS and MetaboLights use the ISA-Tab metadata tracking framework (Sansone *et al.* 2012) and emerging data publication platforms such *Nature Publishing Group’s Scientific Data* and BioMedCentral/BGI’s GigaScience already represent experimental information using the ISA framework.

In September 2012, the 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. We will coordinate our effort with NIH.



As the editor of the *Metabolomics* journal states in his editorial “*Metabolomics* [the journal] is already on record in saying that it wishes studies that it publishes and data therein to be as MSI compliant as possible (Goodacre 2010)”. Whilst established procedures are not yet in place for metabolite data upload within an MSI compliant decentralised framework, COSMOS is encouraging researchers to deposit their data to one of the above repositories; MetaboLights or Metabolomics Workbench. Meanwhile, MSI compliant standards of data annotation, reporting, management and flow are being developed, promoted and entrenched so that those can be shared routinely and re-used effectively.

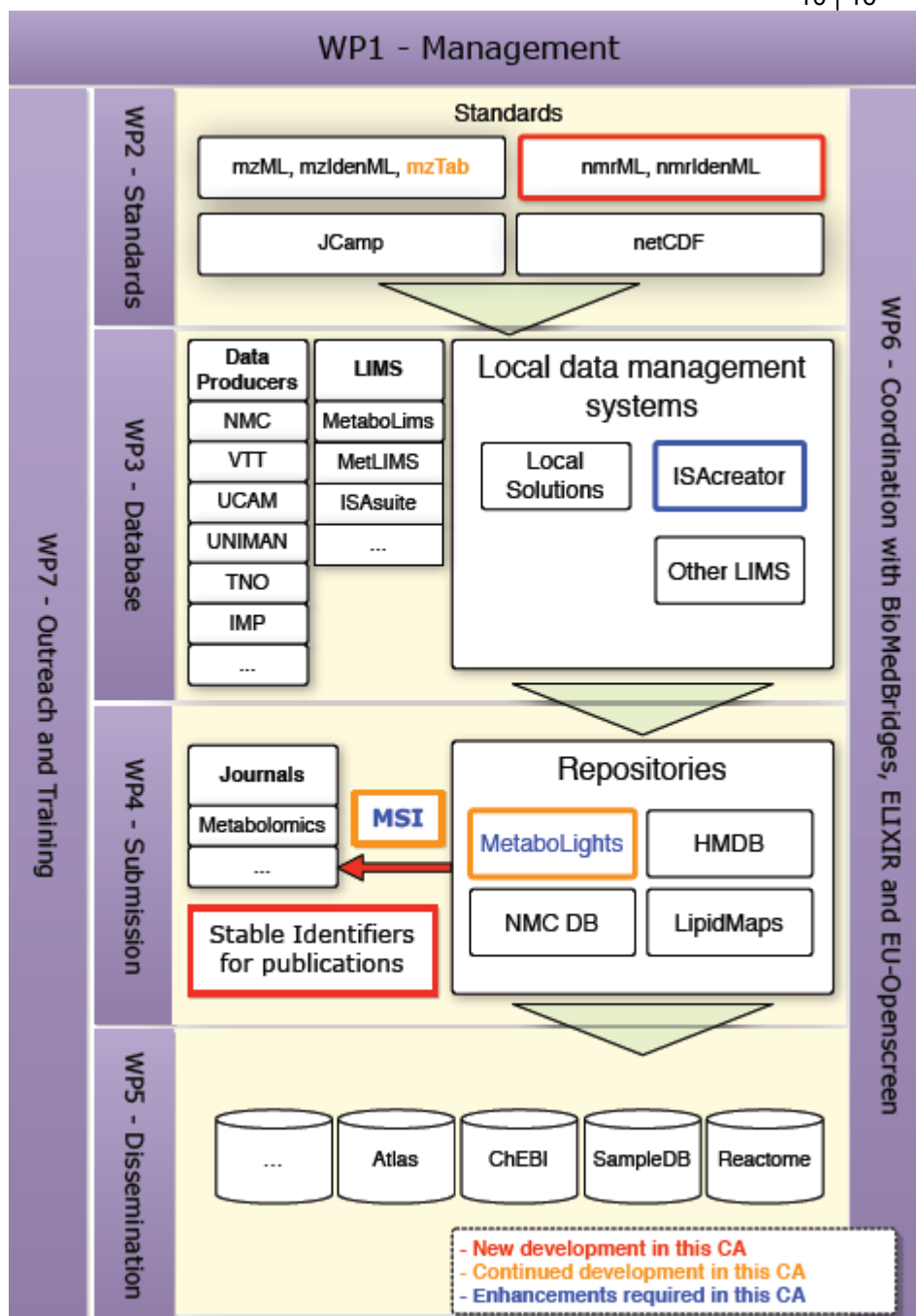
In time this and the other activities discussed above will establish objective methods by which metabolomics data and accompanying metadata may be associated with publications seamlessly, thus allowing easy access. These data will be tagged with a unique identifier generated by COSMOS that will identify and be associated with data and paper. Once these are in place the research community capacity to manage and analyse metabolomics data will be strengthened resulting in higher quality science with larger sharing and better value for research funders. Table 2 lists relevant URLs related to the work described above.



**Table 2:** List of relevant URLs.

COSMO consortium	<a href="http://www.cosmos-fp7.eu/">http://www.cosmos-fp7.eu/</a>
MetaboLights	<a href="http://www.ebi.ac.uk/metabolights/">http://www.ebi.ac.uk/metabolights/</a>
Metabolomics Workbench	<a href="http://www.metabolomicsworkbench.org">http://www.metabolomicsworkbench.org</a>
<i>Metabolomics</i> journal	<a href="http://link.springer.com/journal/11306">http://link.springer.com/journal/11306</a>
Metabolomics Standards Initiative (MSI)	<a href="http://msi-workgroups.sourceforge.net/">http://msi-workgroups.sourceforge.net/</a>
Nature Scientific Data	<a href="http://www.nature.com/scientificdata/">http://www.nature.com/scientificdata/</a>
GigaScience	<a href="http://www.gigasciencejournal.com/">http://www.gigasciencejournal.com/</a>
Isatools	<a href="http://isa-tools.org/">http://isa-tools.org/</a>
National Institutes of Health	<a href="http://www.nih.gov/">http://www.nih.gov/</a>

The work plan of the COSMOS coordination action follows the natural workflow of how Metabolomics data is generated, captured, stored and disseminated in biomedical and life science studies and focuses on developing policies, open standards and open software to ensure the maximum ease of use of Metabolomics data in life sciences and biomedical e-infrastructures, see Figure 2.



**Figure 2:** General COSMOS Proposal Structure.



### 3.3 Next steps

In time this and the other activities discussed above will certainly establish objective methods by which metadata and metabolomics data may be associated with publications seamlessly, thus allowing easy access. These data will be tagged with a unique identifier generated by COSMOS that will identify and be associated with data and paper. Once these are in place the research community capacity to manage and analyse metabolomics data will be strengthened resulting in higher quality science and better value for research funders. Table 2 lists relevant URLs related to the work described above.

**Table 1:** List of relevant URLs.

COSMO consortium	<a href="http://www.cosmos-fp7.eu/">http://www.cosmos-fp7.eu/</a>
<i>Metabolomics</i> journal	<a href="http://link.springer.com/journal/11306">http://link.springer.com/journal/11306</a>
Metabolomics Standards Initiative (MSI)	<a href="http://msi-workgroups.sourceforge.net/">http://msi-workgroups.sourceforge.net/</a>

## 4 Publications

N/A

## 5 Delivery and schedule

The delivery is delayed: ☐ Yes ☒ No

## 6 Adjustments made

This work would be continuously change and update and requirement change by time



## 7 Efforts for this deliverable

Institute	Person-months (PM)		Period
	actual	estimated	
9: UNIMAN	5	12	15
2: LU/NMC	1		
1: EMBL-EBI	2		
3: MRC	0.6		
8: MPG	1		
6: VTT	0.81		
Total	10.41		

## 8 References

- Fiehn O., Robertson D., Griffin J., van der Werf M., Nikolau B., Morrison N., Sumner L.W., Goodacre R., Hardy N.W., Taylor C., Fostel J., Kristal B., Kaddurah-Daouk R., Mendes P., van Ommen B., Lindon J.C., Sansone S.A. (2007). The metabolomics standards initiative (MSI). *Metabolomics* 3, 175–178.
- Fiehn O., Sumner L.W., Rhee S.Y., Ward J., Dickerson J., Lange B.M., et al. (2007b). Minimum reporting standards for plant biology context information in metabolomic studies. *Metabolomics* 3, 195-201.
- Goodacre R. (2014). Water, water, every where, but rarely any drop to drink. Editorial, *Metabolomics* 10, 5-7.
- Goodacre R. (2010). An overflow of... what else but metabolism! *Metabolomics*, 6, 1-2.



- Goodacre R., Broadhurst D., Smilde A., Kristal B.S., Baker J.D., Beger R., et al. (2007). Proposed minimum reporting standards for data analysis in metabolomics. *Metabolomics*, 3, 231-241.
- Griffin J.L., Nicholls A.W., Daykin C., Heald S., Keun H., Schuppe-Koistinen I., et al. (2007). Standard reporting requirements for biological samples in metabolomics experiments: Mammalian/in vivo experiments. *Metabolomics* 3, 179-188.
- Hardy N.W. and Taylor C.F. (2007). A roadmap for the establishment of standard data exchange structures for metabolomics. *Metabolomics* 3, 243-248.
- Morrison N., Bearden D., Bundy J.G., Collette T., Currie F., Davey M.P., et al. (2007). Standard reporting requirements for biological samples in metabolomics experiments: Environmental context. *Metabolomics*, 3, 203-210.
- Rubtsov D.V., Jenkins H., Ludwig C., Easton J., Viant M. R., Gunther U., et al. (2007). Proposed reporting requirements for the description of NMR-based metabolomics experiments. *Metabolomics*, 3, 223-229.
- Sansone S.-A., Schober D., Atherton H. J., Fiehn O., Jenkins H., Rocca-Serra P., et al. (2007). Metabolomics standards initiative – ontology working group: Work in progress. *Metabolomics* 3, 249-256.
- Sansone et al <http://www.nature.com/nbt/journal/v25/n8/full/nbt0807-846b.html>
- Sansone et al. <http://www.nature.com/ng/journal/v44/n2/full/ng.1054.html>
- Sumner L.W., Amberg A., Barrett D., Beger R., Beale M.H., Daykin C., et al. (2007). Proposed minimum reporting standards for chemical analysis. *Metabolomics*, 3, 211-221.
- van der Werf M.J., Takors R., Smedsgaard J., Nielsen J., Ferenci T., Portais J.C., et al. (2007). Standard reporting requirements for biological samples in metabolomics experiments: Microbial and in vitro biology experiments. *Metabolomics*, 3, 189-194.



## 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	W P 4	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. Define COSMOS metadata format, as formally agreed by the members											



### 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