Deliverable D4.2

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 metadata format 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 July 2013	
WP leader:	Roy Goodacre	UNIMAN

Contributing partner(s):	Elon Correa (WP4), Jan Hummel, (WP3) Theo Reamers (WP5), Jules Giffin, Philippe Rocca-Serra, Steffen Neumann (WP2), Matej Oresic, Reza Salek (WP1), Roy Goodacre (WP4),

Authors: Elon Correa, Jan Hummel, Theo Reamers, Reza Salek, Roy Goodacre

Contents

1 Executive summary

This deliverable aims to describe minimum set of agreeable COSMOS metadata format definition, as formally agreed by the members of the COSMOS consortium to ensure correct and proper use, reporting and interpretation of the data by its owners and users. To achieve this common schema, a number of characteristics, or attributes, extracted from or associated with the original dataset are currently being carefully designed, proposed and agreed by all COSMOS participants, stakeholders, vendors and publishers.

2 Project objectives

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

No.	Objective	Yes	No
1	Define COSMOS metadata format, as formally agreed by the members of the COSMOS consortium	X	

3 Detailed report on the deliverable

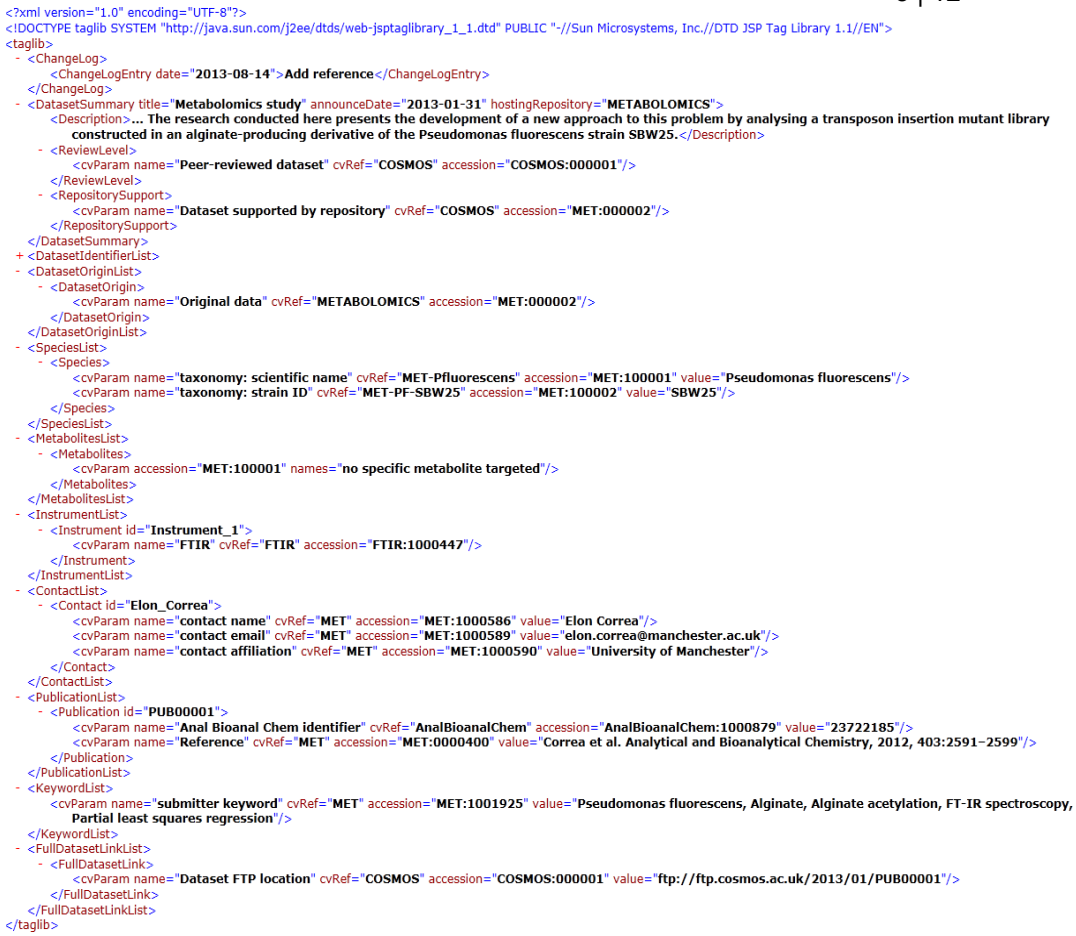
3.1 Background

COSMOS will automatically generate structured information “metadata” to annotate succinctly any data deposited by collaborating partners. The metadata will contain sufficient information to (I) unambiguously identify the experiment, (ii) briefly describe study design, submitter and reference database location where the data is stored and (iii) make data easily retrievable, possibly via direct links. For data that involve experimental setup the metadata should also include enough information to allow faithful reproduction of experiments and results. Given the diversity of databases and datasets that COSMOS will access, metadata

standards specification is a long and democratic process based on community-agreed decisions. Therefore, the structure described here is not considered final and is likely to evolve throughout the lifetime of project as new collaborators, databases and publishers requirements change for data annotation. Users, editors and reviewers will benefit from automated data consistency checks implemented by the repositories, as well as from automatic metadata capture and transmission after data deposition.

3.2 Description of Work

The COSMOS metadata format definition is tightly linked to data format decisions made in WP2 and implemented in WP3 and WP5. However, we initially propose the following general minimum information metadata definition, which contains basic textual information to describe, identify and locate data. Figure 1 displays the XML document schema of a general COSMOS metadata file format (see also Appendices 1 & 2). The document starts with an entry reference date and entry record followed by the title of the data, its submission date, the identification of the hosting repository and a brief description of the data. Other information such as COSMOS accession number, study type, contact ID, related publication and database location are also included. The desired metadata 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. For example; ISAconverter will be extended in order to ensure conversion of ISA investigation file to the proposed “xml format”. The ISA investigation file is a declarative file providing key information about experimental design, experimental variable, methods, contacts and bibliographic information. The ISA investigation file can also be further enriched to keep track of funding information and support. It therefore is already a vehicle for most of the metadata that can be found in MetaboLights.



```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE taglib SYSTEM "http://java.sun.com/j2ee/dtds/web-jsptaglibrary_1_1.dtd" PUBLIC "-//Sun Microsystems, Inc.//DTD JSP Tag Library 1.1//EN">
<taglib>
  - <ChangeLog>
      <ChangeLogEntry date="2013-08-14">Add reference</ChangeLogEntry>
    </ChangeLog>
  - <DatasetSummary title="Metabolomics study" announceDate="2013-01-31" hostingRepository="METABOLOMICS">
      <Description>... The research conducted here presents the development of a new approach to this problem by analysing a transposon insertion mutant library
        constructed in an alginate-producing derivative of the Pseudomonas fluorescens strain SBW25.</Description>
    - <ReviewLevel>
        <cvParam name="Peer-reviewed dataset" cvRef="COSMOS" accession="COSMOS:000001"/>
      </ReviewLevel>
    - <RepositorySupport>
        <cvParam name="Dataset supported by repository" cvRef="COSMOS" accession="MET:000002"/>
      </RepositorySupport>
    </DatasetSummary>
  + <DatasetIdentifierList>
  - <DatasetOriginList>
    - <DatasetOrigin>
        <cvParam name="Original data" cvRef="METABOLOMICS" accession="MET:000002"/>
      </DatasetOrigin>
    </DatasetOriginList>
  - <SpeciesList>
    - <Species>
        <cvParam name="taxonomy: scientific name" cvRef="MET-Pfluorescens" accession="MET:100001" value="Pseudomonas fluorescens"/>
        <cvParam name="taxonomy: strain ID" cvRef="MET-PF-SBW25" accession="MET:100002" value="SBW25"/>
      </Species>
    </SpeciesList>
  - <MetabolitesList>
    - <Metabolites>
        <cvParam accession="MET:100001" names="no specific metabolite targeted"/>
      </Metabolites>
    </MetabolitesList>
  - <InstrumentList>
    - <Instrument id="Instrument_1">
        <cvParam name="FTIR" cvRef="FTIR" accession="FTIR:1000447"/>
      </Instrument>
    </InstrumentList>
  - <ContactList>
    - <Contact id="Elon_Correa">
        <cvParam name="contact name" cvRef="MET" accession="MET:1000586" value="Elon Correa"/>
        <cvParam name="contact email" cvRef="MET" accession="MET:1000589" value="elon.correa@manchester.ac.uk"/>
        <cvParam name="contact affiliation" cvRef="MET" accession="MET:1000590" value="University of Manchester"/>
      </Contact>
    </ContactList>
  - <PublicationList>
    - <Publication id="PUB00001">
        <cvParam name="Anal Bioanal Chem identifier" cvRef="AnalBioanalChem" accession="AnalBioanalChem:1000879" value="23722185"/>
        <cvParam name="Reference" cvRef="MET" accession="MET:0000400" value="Correa et al. Analytical and Bioanalytical Chemistry, 2012, 403:2591–2599"/>
      </Publication>
    </PublicationList>
  - <KeywordList>
      <cvParam name="submitter keyword" cvRef="MET" accession="MET:1001925" value="Pseudomonas fluorescens, Alginate, Alginate acetylation, FT-IR spectroscopy,
        Partial least squares regression"/>
    </KeywordList>
  - <FullDatasetLinkList>
    - <FullDatasetLink>
        <cvParam name="Dataset FTP location" cvRef="COSMOS" accession="COSMOS:000001" value="ftp://ftp.cosmos.ac.uk/2013/01/PUB00001"/>
      </FullDatasetLink>
    </FullDatasetLinkList>
</taglib>
```

Figure 1: XML schema of a general COSMOS metadata file.

3.3 Next steps

The final metadata format will be heavily influenced by decisions made in WP2 and implemented in WP3 and WP5. It will also be supported by a communally developed set of definitions for metadata capture and transmission after data deposition. We are liaising with all parties interested and striving to make sure that COSMOS metadata format definitions are fully supported by vendors and publishers, who require deposition upon publication. In addition, we will need to continue the development of standards for metadata to maximize the utility of this resource.

4 Publications

N/A

5 Delivery and schedule

The delivery is delayed: ☐ Yes ☑ No

6 Adjustments made

This work, might change and updated as the requirements change by time in accordance to data work flow schema and negotiation with data repositories.

7 Efforts for this deliverable

Institute	Person-months (PM)		Period
	actual	estimated	9
9: UNIMAN	1		
2:LU/NMC	1 (in Kind)		
1:EMBL-EBI	0.5		
3:MRC	1		
6:VTT	0.26		
314:UOXF	0.5 (in kind)		
8:MPG	0.5		
Total	5.77	9	

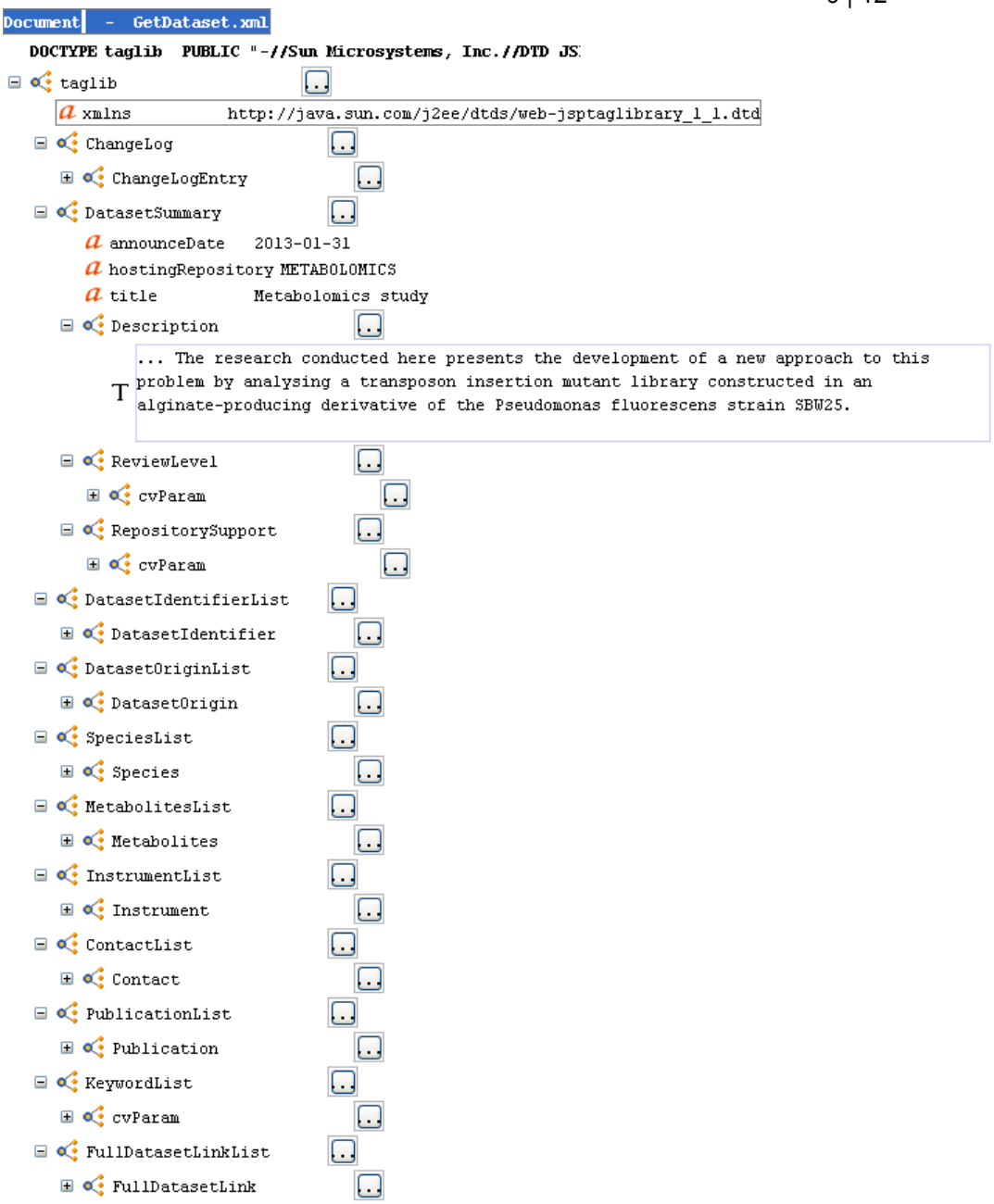
Appendices

Appendix 1: XML schema of a general COSMOS metadata file

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE taglib PUBLIC "-//Sun Microsystems, Inc.//DTD JSP Tag
Library 1.1//EN" "http://java.sun.com/j2ee/dtds/web-
jsptaglibrary_1_1.dtd">
<taglib>
  <ChangeLog>
    <ChangeLogEntry date="2013-08-14">Add
reference</ChangeLogEntry>
  </ChangeLog>
  <DatasetSummary title="Metabolomics study"
hostingRepository="METABOLOMICS" announceDate="2013-01-31">
    <Description>... The research conducted here presents the
development of a new approach to this problem by analysing a
transposon insertion mutant library constructed in an alginate-
producing derivative of the Pseudomonas fluorescens strain
SBW25.</Description>
    <ReviewLevel>
      <cvParam accession="COSMOS:000001" cvRef="COSMOS"
name="Peer-reviewed dataset"/>
    </ReviewLevel>
    <RepositorySupport>
      <cvParam accession="MET:000002" cvRef="COSMOS" name="Dataset
supported by repository"/>
    </RepositorySupport>
  </DatasetSummary>
  <DatasetIdentifierList>
    <DatasetIdentifier>
      <cvParam accession="FTIR:0000001" cvRef="FTIR"
value="FTIR01" name="COSMOS accession number"/>
    </DatasetIdentifier>
  </DatasetIdentifierList>
  <DatasetOriginList>
    <DatasetOrigin>
      <cvParam accession="MET:000002" cvRef="METABOLOMICS"
name="Original data"/>
    </DatasetOrigin>
  </DatasetOriginList>
  <SpeciesList>
    <Species>
      <cvParam accession="MET:100001" cvRef="MET-Pfluorescens"
value="Pseudomonas fluorescens" name="taxonomy: scientific name"/>
      <cvParam accession="MET:100002" cvRef="MET-PF-SBW25"
value="SBW25" name="taxonomy: strain ID"/>
    </Species>
  </SpeciesList>
  <MetabolitesList>
    <Metabolites>
      <cvParam accession="MET:100001" names="no specific
metabolite targeted"/>
    </Metabolites>
  </MetabolitesList>
  <InstrumentList>
    <Instrument id="Instrument_1">
```

```
      <cvParam accession="FTIR:1000447" cvRef="FTIR" name="FTIR"/>
    </Instrument>
  </InstrumentList>
  <ContactList>
    <Contact id="Elon_Correa">
      <cvParam accession="MET:1000586" cvRef="MET" value="Elon
Correa" name="contact name"/>
      <cvParam accession="MET:1000589" cvRef="MET"
value="elon.correa@manchester.ac.uk" name="contact email"/>
      <cvParam accession="MET:1000590" cvRef="MET"
value="University of Manchester" name="contact affiliation"/>
    </Contact>
  </ContactList>
  <PublicationList>
    <Publication id="PUB00001">
      <cvParam accession="AnalBioanalChem:1000879"
cvRef="AnalBioanalChem" value="23722185" name="Anal Bioanal Chem
identifier"/>
      <cvParam accession="MET:0000400" cvRef="MET" value="Correa
et al. Analytical and Bioanalytical Chemistry, 2012, 403:2591-
2599" name="Reference"/>
    </Publication>
  </PublicationList>
  <KeywordList>
    <cvParam accession="MET:1001925" cvRef="MET"
value="Pseudomonas fluorescens, Alginate, Alginate acetylation,
FT-IR spectroscopy, Partial least squares regression"
name="submitter keyword"/>
  </KeywordList>
  <FullDatasetLinkList>
    <FullDatasetLink>
      <cvParam accession="COSMOS:000001" cvRef="COSMOS"
value="ftp://ftp.cosmos.ac.uk/2013/01/PUB00001" name="Dataset FTP
location"/>
    </FullDatasetLink>
  </FullDatasetLinkList>
</taglib>
```

Appendix 2: A snapshot of a graphical view of XML schema of a general COSMOS metadata file



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	WP 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 of the COSMOS consortium

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