Deliverable D2.5

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:	Real converters, parsers & validators for NMR-ML	
WP No.	2	
Lead Beneficiary:	11. IPB	
WP Title	Standards Development	
Contractual delivery date:	01 10 2014	
Actual delivery date:	01 10 2014	
WP leader:	Steffen Neumann	IPB
Contributing partner(s):	11. IPB, Michael Wilson from Wishart Lab, University of Alberta, Edmonton Canada, 1.EMBL-EBI, 12 UB2, 13 UBHam (in kind contribution), 14 UOXF	

Authors: Authors: Authors: Daniel Schober, Michael Wilson, Annick Moing, Daniel Jacob, Jie Hao, Tim Ebbels, Reza Salekand Steffen Neumann

Contents

1 Executive summary

For this deliverable D 2.5 we have coordinated efforts from multiple international groups who are developing tools and parsers for the nmrML format. In particular, we here deliver automatic converters that read in proprietary vendor raw data files (Bruker and Varian/Agilent) and generate schema compliant nmrML XML files either manually or in a high-throughput batch mode. These parsers and converters are available for multiple programming languages (JAVA and Python) and can be deployed as web applications, as part of existing software pipelines or as standalone command line tools. We also deliver parser extensions for different established software frameworks such as R and Matlab based packages (e.g. Batman[1]and rNMR[2]), which allow for reading in nmrML files and make their content amenable to statistical analysis. We also had interest from the proprietary Chenomx NMR suite developers to support the format at a later stage.

There is an active developer community, and we expect the development to continue in the future and also beyond COSMOS.

2 Project objectives

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

No.	Objective	Yes	No
1	Deliver software that converts the major proprietary vendor NMR formats into the open nmrML format	X	
2	Deliver parsers that read the open nmrML format and makes its content accessible to open 3rd party processing tools	X	
3	Deliver software that validates existing nmrML files according to quality schemes defined in Minimal Information checklists	X	

[1] Hao, J., Astle, W., De Iorio, M., & Ebbels, T. M. (2012). BATMAN--an R package for the automated quantification of metabolites from nuclear magnetic resonance spectra using a Bayesian model. *Bioinformatics, 28*(15), 2088-2090, doi:10.1093/bioinformatics/bts308.
[2] Lewis, I. A., Schommer, S. C., & Markley, J. L. (2009). rNMR: open source software for identifying and quantifying metabolites in NMR spectra. *Magn Reson Chem, 47 Suppl 1*, S123-126, doi:10.1002/mrc.2526.

3 Detailed report on the deliverable

3.1 Background

NMR is an important analytical method in metabolomics experiments. Currently existing standard data formats such as the JCAMP family have several drawbacks, especially in metabolomics applications. One problem is that there is no semantic validation of JCAMP-DX files. In deliverable D2.4, we introduced the new open nmrML data format specification to capture and freely exchange NMR raw data. To actually use nmrML format with NMR data, we need parsers that convert the vendor file formats into nmrML.

The dominant instrument vendors i.e. Bruker, Varian/Agilent and JEOL, typically provide the instrument software to process the vendor specific data. But alternative data analysis software needs to put considerable efforts into reading and writing these specific vendor formats, this applies both to commercial software such as NmrPipe, MestReNova (Mnova) or Chenomx NMR Suite, but even more so to community developed open source efforts such as the Batman R package, rNMR or Metaboquant[3] (Matlab-based).

Fig. 1. provides an overview of the different parsers and converters tackled in this deliverable and how they contribute to the COSMOS nmr data information flow.

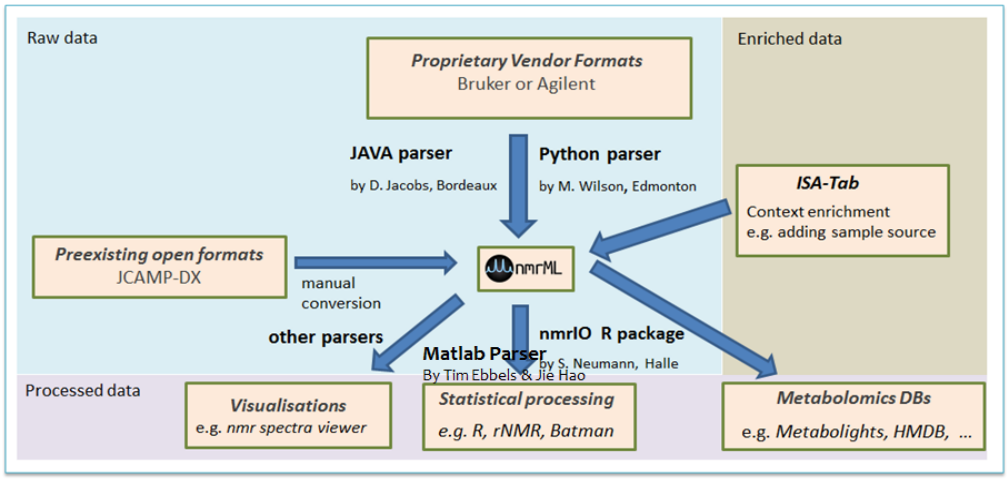

Figure 1: Illustration of NMR data management facilitation by means of the common nmrML standard

[3] Wolfram Gronwald, Matthias Klein and Peter Oefner (2013), MetaboQuant: A Tool Combining Individual Peak Calibration and Outlier Detection for Accurate Quantification from NMR Spectra

Besides the parsers, it is important to develop tools to ensure data quality. We deliver a semantic validator and corresponding webservice, which checks the quality of the generated NMR data files in a multilayered approach, i.e. ensuring that the data is syntactically well formatted, adheres to the nmrML.xsd schema, and is sufficiently detailed with respect to data content and CV annotations. Semantic validation exploits rules that set constraints on certain XML positions, i.e. which CV terms are allowed at a certain XML location. Such checks (see Fig 3) can enforce aspects of minimal information requirements, e.g. from the Core Information for Metabolomics Reporting[4] (CIMR) or given journal policies.

3.2 Description of Work

The work on nmrML was continued over the last year. We had regular teleconferences with a prepared agenda and minutes taken by the participants. A workshop was held in Edmonton (Canada), which was also attended by representatives from Chenomx, one of the leading commercial NMR software companies.

3.2.1 Development of Vendor to nrmML converters

Java based converter

Based on both nmrML.xsd (XML Schema Definition) and CV params (such as ontologies nmrCV, UO, CHEBI ...), a converter written in Java was developed that automatically generates nmrML files, from raw files of the major NMR vendors. The choice of Java was guided by i) the JAXB framework (Java Architecture for XML Binding), ii) its OS-platform independence and iii) strengthened by the existence of a useful java library (i.e nmr-fid-tool) for further processing and visualisation of the resulting nmrML data.

As nmrML intents to gather and integrate several types of data and corresponding metadata in a single file, it is necessary to process each data source separately.

[4] Denis V. Rubtsov *et al.* (2007), Proposed minimum reporting standards for the description of NMR-based metabolomics experiments

Thus, two command tools were developed. The first one, *nmrMLcreate* allows to create a new nmrML file, based on available Bruker, Varian/Agilent or Jeol raw files. The second one, *nmrMLadd*, acts as a wrapper, allowing to add and fill in additional sections corresponding to the data levels, including the data processing step. (cf Figure 2).

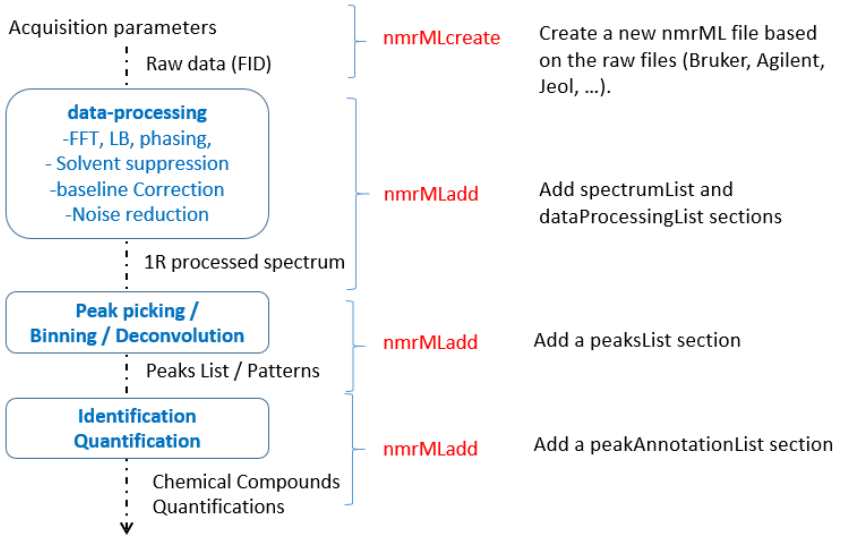

Figure 2: The data workflow related to nmrML: from the raw data up to the final annotation step, nmrML files can be updated by adding a section to the corresponding step.

To make this converter usable without a local installation it is implemented as a lightweight and easy to access web service, for which we also generated tutorial videos.

Python based converter

A python based parser that exploits parameter mappings is available as software code in the nmrML Git developer repository at github.com/nmrML/nmrML/tree/master/tools/Parser_and_Converters/python/pynmrml , including the documentation on installation and usage.

3.2.2 nmrML data validator

We have implemented a CV-aware semantic validator, leveraging on existing OpenMS software and external customizable mapping files containing the concrete verification rules to check nmrML XML instances for semantic errors and completeness. This validation scheme is visualized in Fig. 3 explaining its components and a few example rules exploitable by it.

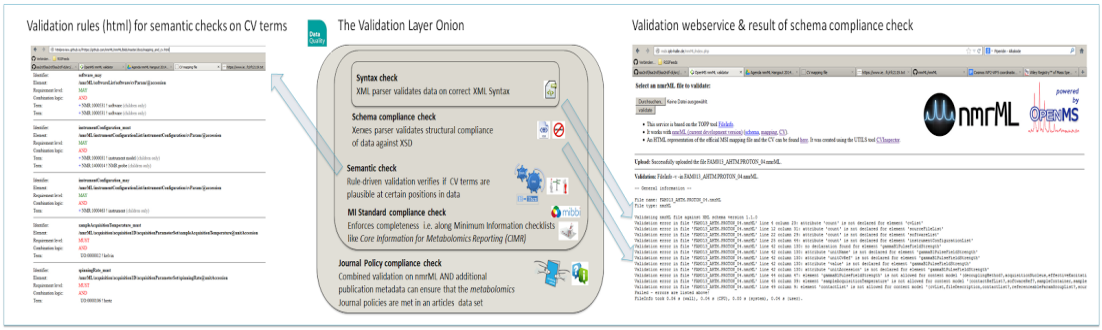

Figure 3: Illustration of onion layered validation of nrmML data files by means of a semantic rule based Validator webservice.

3.2.3 nmrML to processing tool and library parsers

R parsers

The R statistics language, and in particular the Bioconductor packages are widely used in the life-sciences. Originally focussing on e.g. gene expression analysis, today packages are available for proteomics and metabolomics tasks as well. With affyIO and mzR, there are parser packages for gene expression data and mass spectrometry data files, which provide a stable I/O API for the actual data analysis tasks. The IPB is developing the nmRIO package, which aims to provide access to NMR data files for high-level NMR analysis software.

The nmRIO package development is hosted inside the main nmrML github repository at github.com/nmrML/nmrML/tree/master/tools/Parser_and_Converters/R/nmRIO

to ensure that the code is in sync with the latest schema and CV updates. A vignette that combines code examples, text and graphical output and the description of the package is included. The submission to either the CRAN or Bioconductor repositories is planned later.

Matlab parser

The first version of Matlab parser is complete and on Github at
github.com/nmrML/nmrML/tree/master/tools/Parser_and_Converters/Matlab
It reads all fields of nmrML parameters, and decodes FID and spectrum array data. Figure 4 below shows a spectrum plot from Matlab parser with example file 'MMBBI_10M12-CE01-1a.nmrML'.

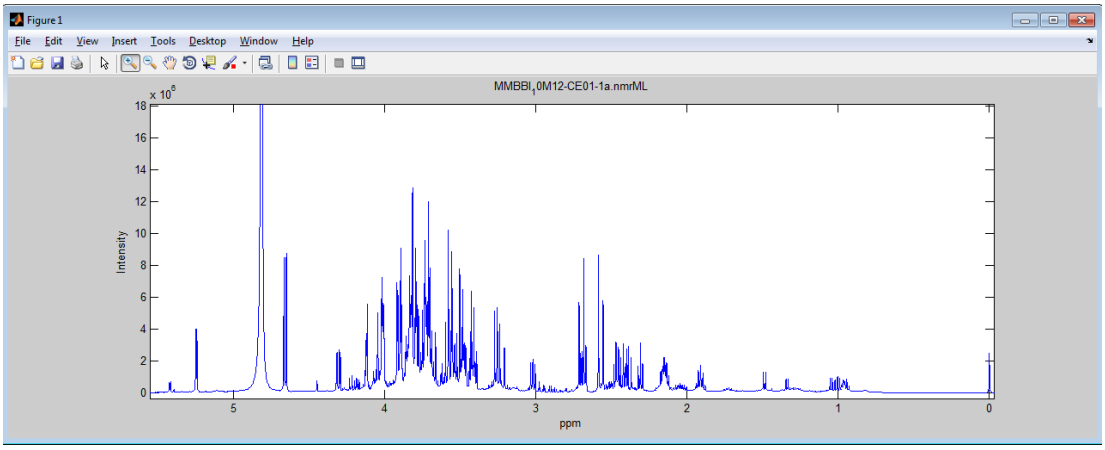

Figure 4: An NMR Spectrum visualized based on nrmML formatted input data.

3.2.4 *Ident- and Quant extensions to nmrML XSD*

As one of the outcomes of the Edmonton workshop, we looked at how the Chemical Markup Language (CML) represents molecules, spectra and assignments, as used by the Wishard Lab HMDB NMR spectrum viewer. We have adopted the molecule description from CML and added peak assignment to our nmrML xsd, based on a data-driven HMDB example (i.e. Propionic acid).

3.2.5 *Tool access and documentation*

All source files are available on the project Github pages, together with accompanying readme files. The operating services are accessible from our nmrml.org website:

GitHub for tool sources: github.com/nmrML/nmrML/tree/master/tools
nmrML parsers and converters: nmrml.org/converter/
nmrML semantic validator: nmrml.org/validator/ https://groups.google.com/forum/

3.3 Next steps

Further testing of the XSD is required with diverse experimental configurations, to ensure sufficient coverage. We must also ensure that the parsers and converters keep up and in sync with eventual XML schema additions and changes, e.g. our latest additions with respect to quantification and compound identification. Rule sets can be specified with different stringency for validation on distinct quality levels. The creation of ISA Tab specifications for easy tabular data entry and minimal reporting requirement enforcement is considered a further next step (D2.6).

4 Publications

- Daniel Schober, Michael Wilson, Daniel Jacob, Annick Moing, Gerhard Mayer, Martin Eisenacher, Reza M Salek and Steffen Neumann: **Ontology Usage in Omics Standards Initiatives: Pros and Cons of Enriching XML Data Formats with Controlled Vocabulary Terms,** *ODLS Proceedings 2014*, Freiburg, www.onto-med.de/obml/ws2014/odls2014report_prefinal.pdf , page 36.
- Second publication in progress

5 Delivery and schedule

The delivery is delayed: ☑ Yes ☑ No

6 Adjustments made

N/A

7 Efforts for this deliverable

Institute	Person-months (PM)		Period
	actual	estimated	24
1: EMBL-EBI	3		
11: IPB	4		
4: IMPERIAL	1		
3:MRC	0.4		
8:MPG	3		
6:VTT	1.66		
10:CIRMMP	4		
12. UB2	2		
14:UOXF	2.24		
Michael Wilson, Wishart Lab	2 (in kind contribution)		
Total	22.3 (2 in kind)	6	6

Appendices

1. N/A

Background information

This deliverable relates to WP2; background information on this WP as originally indicated in the description of work (DoW) is included below.

WP2 Title: Standards Development
Lead: Steffen Neumann, IPB
Participants: EBI-EMBL, LU-NMC, MRC, IMPERIAL, TNO and VTT

This work package will deliver the exchange formats and terminological artifacts needed to describe, exchange and query both the metabolomics data and the contextual information ('experimental metadata' — e.g., provenance of study materials, technology and measurement types, sample-to-data relationships). We will ensure that these standards are widely accepted and used by involving all major global players in the development process. The consortium represented by COSMOS already contains the majority of players in Metabolomics in Europe and other global players in the field have provided letters of support. Those and others will be invited both the work meetings as well as the regular stakeholder meetings. As the open standards developed here are supported by open source tools, they can be easily put to work which will aid adoption.

Work package number	**WP2**	**Start date or starting event:**										Month 1		
Work package title		Standards Development												
Activity Type		COORD												
Participant number	1: EMBL/EBI	2: LU/NMC	3:MRC	4: Imperial	5: TNO	6: VTT	7:UB	8:MPG	9:UNIMAN	10:CIRMMP	11:IPB	12:UB2	13:UBHAM	14:UOXF
Person-months per participant	12	4	2	3	1	4	2	6	2	6	16	6	4	6

Objectives

1. We will develop and maintain exchange formats for raw data and processed information (identification, quantification), building on

experience from standards development within the Proteomics Standards Initiative (PSI). We will develop the missing open standard NMR Markup Language (NMR-ML) for capturing and disseminating Nuclear Magnetic Resonance spectroscopy data in metabolomics. This is urgently needed as long-term archival format if metabolomic databases are to capture all the formats of metabolomic data, as well as supporting developments in cheminformatics and structural biology. For mass spectrometry, we will work with the PSI to extend existing exchange standards to technologies used in metabolomics, e.g. gas chromatography, imaging mass spectrometry and the identification tools and databases.

2. In addition to the raw data formats, we will need to continue the development of standards for experimental metadata and results, independent of the analytical technologies. We will review, maintain and, where needed, extend reporting requirements and terminological artefacts developed by Metabolomics Standards Initiative (MSI). We need to represent quantification options in MS and NMR, and the semantics of data matrices used to summarize experimental results, key information which often is only available in PDF tables associated to manuscripts. As research in biomedical and life sciences is increasingly moving towards multi-omics studies, metabolomics must not be an island. The 'Investigation/Study/Assay' ISA-Tab format was developed to represent experimental metadata independently from the assay technology used. We will use ISA-Tab to standardize metabolomics reporting requirements and terminologies through customized configurations.
3. Finally, we will explore semantic web standards that facilitate linked open data (LOD) throughout the biomedical and life science realms, and demonstrate their use for metabolomics data. While the technical standards already exist, we will need to develop the "inventory" of terms and concepts required to express facts about metabolomics, capturing the data to characterize studies and digital objects in metabolomics to facilitate the data flow in biomedical e-infrastructures.

Description of work and role of participants

Task 1: Development of data exchange formats for Metabolomics data To capture and exchange raw- and processed mass spectrometry data, we will extend existing open standard (such as mzML, mzIdentML and mzQuantML developed by the PSI) to meet the requirements specific to metabolomics experiments. The MPG will add features missing to handle GC/MS, and the IPB work to represent metabolite identification and -quantitation. MRC will work to promote imzML into an MSI approved exchange format for MS based imaging (MALDI, DESI, SIMS). A new data exchange standard is required for the exchange of NMR spectroscopy based metabolomics data. Building on the excellent

experience with XML based formats we will develop the NMR-ML format, a corresponding controlled vocabulary and coordinate the implementation of parsers and tools for validation. Instrument vendors and authors of NMR tools and -databases will be invited to the initiative. The IPB will contribute their expertise from mzML, CIRMMP, including the University of Florence as a third party of CIRMMP, EBI, UBHam and MRC are already involved in discussion with David Wishart from HMDB about NMR-ML.

Task 2: Common representation for Minimum Information Standards for Metabolomics In this WP, we will build on the BioSharing and the ISA-Tab efforts to harmonize representation of the metadata recommendations with other -omics communities, and use automated tests to ensure the interoperability of the metadata between the involved data producers, -consumers and -repositories. The EBI, IPB and MRC will be working with the UOXF to create both core and extended configurations (specific to the research discipline and technologies) suitable for metabolomics, in compliance with the annotation manual created in WP4. This will include a component to report stable isotope labelling and its detection by both mass spectrometry and NMR spectroscopy, required by the metabolomics community carrying out fluxomic studies.

Task 3: Enabling the integration of metabolomics data into large e-science infrastructures. The technologies around the Resource Description Framework (RDF) are used to represent and link the information stored in databases by interconnecting them, relying on a strict semantics for distributed data. Several ontologies of terms and concepts exist for the biological and biomedical domain. In this task we will collect and if necessary extend this inventory to describe metabolomics facts with contributions to existing vocabulary efforts. IPB and UOXF will contribute to e.g. the Ontology for Biomedical Investigations (OBI) and PSI-MS to ensure complete coverage of the key areas of metabolomics technology as a community efforts, leveraging existing, proven infrastructures, in a ‘good citizenship’ frame of mind to avoid duplication of effort. To connect different sources of data and knowledge, the “Semantic Web for Health Care and Life Sciences Interest Group” (HCLSIG) has started work to represent ISA-Tab metadata as RDF, in compliance with the recommendations of the international Linked Data community (http://linkeddata.org), which will allow to expose any ISA-Tab data set to the semantic web. To demonstrate the feasibility, we will create exemplary semantic query endpoints. The EBI, MPG and IPB will augment their MetaboLights, GMD and MassBank databases. We will also jointly create metabolomics-specific guideline documents for semantic annotation, to maximise the interoperability and link ability of e-resources in the biomedical and life sciences.

Data standards will be described by a set of documents, including 1) the description of use cases, architecture design, and the detailed description of the standard 2) the machine readable standard definition, required for the automatic validation of the content expressed in a standard format 3) several example documents covering the use cases and finally 4) one or more reference implementations. These prototype

	implementations help to 1) identify shortcomings of the standard definition during the design phase that only crop up during the implementation and practical use, and 2) speed up the adoption in the bioinformatics community that develops metabolomics related software. The standards defining documents will be discussed during regular phone conferences and at the regular meetings, and developed using open and public repositories. Before they are adopted as MSI standards, they will be sent out to the wider community for a public discussion period. In WP4 we will ensure that international societies and journals make recommendations to use the standards defined in WP2.
	Deliverables

No.	Name	Due month
D2.1	Completion of GC-MS for mzML	6
D2.2	Data exchange format for metabolite identification	12
D2.3	Data exchange format for metabolite quantitation	12
D2.4	Definition of NMR-ML Schema, initial MSI-NMR ontology, example files	12
D2.5	Real data, Converters, Validators and Parsers for NMR-ML	24
D2.6	Collection of ISA configurations for metabolomics studies	27
D2.7	Test infrastructure for the validation of ISA datasets	36
D2.8	Guideline document on RDF and SPARQL for metabolomics resources	24
D2.9	Public availability of query endpoints for linked data from EBI, MPG, IPB	36