Deliverable D2.8

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:	Guideline Document on RDF and SPARQL for metabolomics resources	
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, 1.EMBL-EBI, 3. MRC, 2. MPG, 14 UOXF	

Authors: Authors: Authors: Daniel Schober, Steffen Neumann, Philippe Rocca-Serra, Alejandra Gonzalez-Beltran, Susanna Sansone

Contents

1 Executive summary

There are a large number of data resources in many areas of life-science, including metabolomics. However, it is usually very difficult -- if not impossible -- to perform distributed analysis and create queries across the data resources.

With semantic web standards that facilitate linked open data (LOD), we demonstrate their use for metabolomics data. While the technical standards (e.g. RDF and virtuoso server) already exist, we will needed to develop the “inventory” of terms and concepts required to express facts about metabolomics. We need to provide agreed-upon terminological descriptors, e.g. to characterize studies and digital objects in metabolomics. Establishing such consensus terminologies will facilitate the data flow in biomedical e-infrastructures.

In a first step, we performed a survey of relevant data resources and existing LOD approaches to create, store and query semantic web data services for metabolomics. In addition to building RDF schemata to describe the LOD data content of established Metabolomics data providers, we implemented several prototype resources, so called SPARQL endpoints, to test the RDF models, data conversions and querying. This culminated into a guideline document describing the current state, some best practices and future requirements for data service providers in metabolomics.

2 Project objectives

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

No.	Objective	Yes	No
1	We will explore semantic web standards that facilitate linked open data (LOD) throughout the biomedical and life science	X	

	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.		

3 Detailed report on the deliverable

3.1 Background

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 semi-formal Subject-Predicate-Object (SPO) triple based RDF model for distributed data (Fig. 1). Several existing controlled vocabularies and ontologies provide canonized terms for the biological and biomedical domain. In this task we collect and if necessary extend this inventory to describe metabolomics data. Where applicable, we re-use and contribute to existing vocabulary efforts. IPB, MPG and UOXF 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 community efforts, leveraging existing, proven infrastructures, in a ‘good citizenship’ frame of mind to avoid duplication of effort. We will however mainly leverage on those artefacts that are in harmony with established semantic web best practices and which will allow to achieve production mode data access and SPARQL querying in a realistic time frame, with simplicity, usability and end user compliance as driving goals.

To demonstrate the feasibility, we create exemplary semantic web query endpoints and will later connect these for distributed integrative querying. The EBI, MPG and IPB will augment their MetaboLights, GMD and MassBank databases with LOD resources.

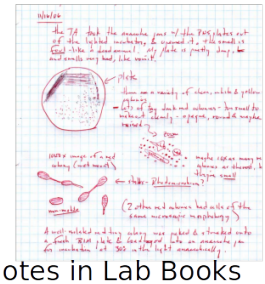
Here, we report on a jointly created metabolomics-specific living guideline document for semantic web data linkage, to describe the current state, some best

practices and future requirements to maximise the interoperability and likability of e-resources in the biomedical and life sciences.

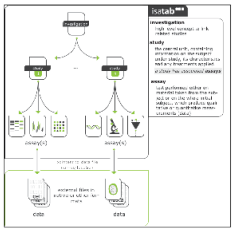
3.2 Description of Work

3.2.1 Description of Use Cases in Metabolomics

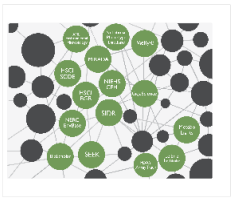
We defined our particular use cases by means of competency questions that we ultimately want to be able to answer by cross resource SPARQL querying. As an established set of technical standards begins to emerge, we need to select the ones most appropriate for our use cases. For this reason, we started to review existing Semantic Web resources in our domain, i.e. the EBIs RDF guideline[1] and the Bio2RDF guidelines[2]. We also reviewed guideline documents by general policy providers like the W3C consortium.



Notes in Lab Books
(information for humans)



Spreadsheets and Tables
(the compromise)



Facts as RDF statements
(information for machines)

Figure 1: Use Case for Metabolomics knowledge representation as RDF statements

We have discussed the use of RDF with the MassBank consortium at the metabolomics conference in Tsuruoka, JP, in June 2014 and at the NORMAN MassBank workshop in Dübendorf, CH in September 2014. The RDF output was designated as one of the future output formats for the whole MassBank consortium.

Another area which deserved review was the tools to be used for making the LOD data accessible over the web. Here, we were mainly guided by three criteria: user

[1] http://www.ebi.ac.uk/rdf/rdf-first-principles
[2] https://github.com/bio2rdf/bio2rdf-scripts/wiki/Bio2RDF-Release-2-ICBO-Tutorials

community size, stability and openness. We leverage on those technologies, which promise an easy future integration of additional emerging relevant endpoints.

3.2.2 Competency Questions for the Metabolomics Use Case

We have collected a set of competency questions, which we want an integrative cross resource SPARQL query engine to be able to answer:

1. Select all MassBank records about certain ChEBI compounds
2. Select all MetaboLights studies which mention a metabolite for which there is a MassBank record
3. Select all compounds from MetaboLights, which are mentioned for Brassicacae species
4. Select all compounds from MetaboLights, used as "insecticide" (CHEBI:24852)
5. Select all MassBank records for molecules that interact with Protein/Enzyme X (using e.g. http://stitch.embl.de/)
6. Select all MassBank records of samples measured in e.g. Germany, Europe or Switzerland (this will require annotation of the records with the gazetteer ontology).

The collection of possible Use Cases also included a review of relevant existing semantic web resources. There are already resources available at the EBI[3], and the bio2rdf project[4] at Carlton University, CA. http://www.cosmos-fp7.eu/

3.2.3 Conversion of MetaboLights Metadata description to RDF using LinkedISA component

In order to expose ISA-Tab coded datasets to the semantic web and the linked open data cloud, the ISA team worked at delivering a converter, the LinkedISA software module, which makes explicit the meaning of ISA tables, the relation between fields (ISA syntactic elements) and their annotations. The work has been

[3] http://www.ebi.ac.uk/about/news/press-releases/RDF-platform
[4] http://bio2rdf.org/

placed in the context of international communities and compliance to best practices. This ranges from recommendations entity identification and RDF creation to ontology development (as coverage gaps need to be addressed). Respectively, UOXF followed recommendations by the international Linked Data community (http://linkeddata.org) and the OBO Foundry. The latter organization was considered as it is an umbrella to several essential biomolecular and model organism semantic representations (Gene Ontology, Phenotype and Trait Ontology, Human Phenotype Ontology, Ontology for Biomedical Investigation, Chemical Entities of Biological Interest), which all share a common semantic framework, therefore, facilitating interoperation and data integration. The LinkedISA implementation however allows including multiple mappings from the ISA syntax to ontologies, in order to support different semantic frameworks.

The LinkedISA conversion component produced by UOXF allows the transformation of an ISA formatted study into an RDF named graph, and offers the ability to carry out further validation checks and automatically augment annotation without user intervention, by taking advantage of the underlying semantic model and a number of Semantic Web Rule Language (SWRL) rules.
The extra layer of validation revealed that about 80% of the experiments stored in MetaboLights can be represented this way. The tests against case-queries, such as verification of study design main features (sample size, factorial, balance) validated the approach and provides requirements for curation tasks and future curation guidelines. In fact, this work identified the need to enforce stricter curation rules and implementation guidelines for a number of common patterns of information (not) found in metabolomics studies.
The extra information automatically added to the named graph during the conversion process allows the following queries to be performed much more efficiently:

1. select all studies with at least 3 samples per study group using targeted metabolite profiling
2. select all studies with a balanced factorial design
3. select all samples and data files from control or untreated groups.

The knowledge gained in this work is readily exploited through the iterative refinement of ISAconfigurations, collection of terminology requests and documentation of coding patterns. In order to fully benefit from the RDF representation, in the future, the linkedISA conversion tool could be integrated to the submission pipeline to aid the curation and validation of MetaboLights submissions.

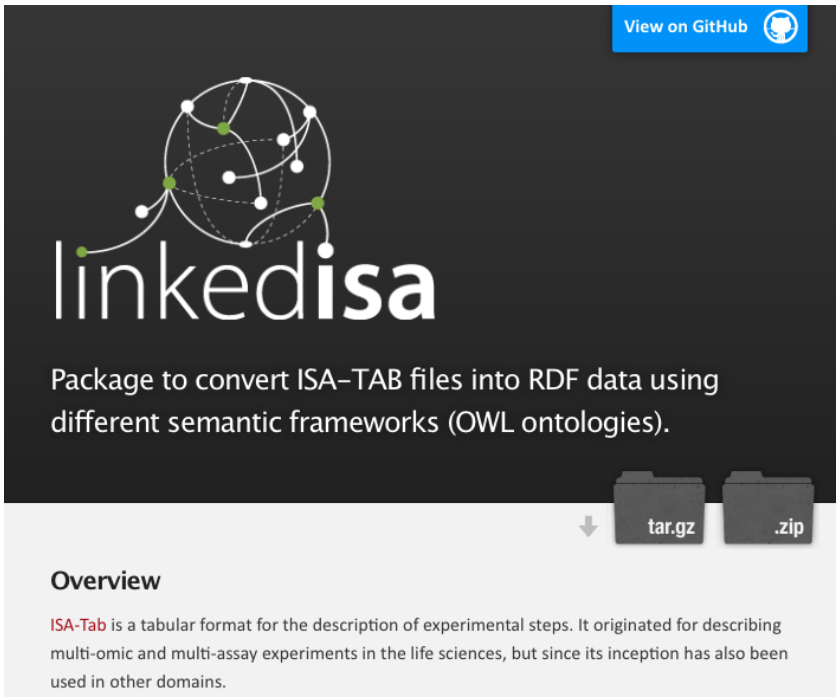


Figure 2: linkedISA website

The existing conversion already enables easy cohort creation. UOXF is currently implementing the conversion from “Metabolite assignment files” (MAF) file to RDF to enable querying from experimental metadata to chemical identities and vice-versa.

Specific gaps in coverage in the semantic resources have been identified and will need addressing. ISA Team, under COSMOS, has been collecting use cases and terms in a series of user meetings (Barcelona Fluxomics Meeting, UK-China meeting at BGI, interaction with the companies Biocrates AG and Bruker Daltonics). It will require community outreach to reach agreement and issue implementation guidelines.

Github:

Converter code ISA-Tab archive to RDF:

http://isa-tools.github.io/linkedISA/

https://github.com/ISA-tools/linkedISA

Web Application: Experimental Metadata Repository RDF based Prototype: http://bii.oerc.ox.ac.uk/browse/

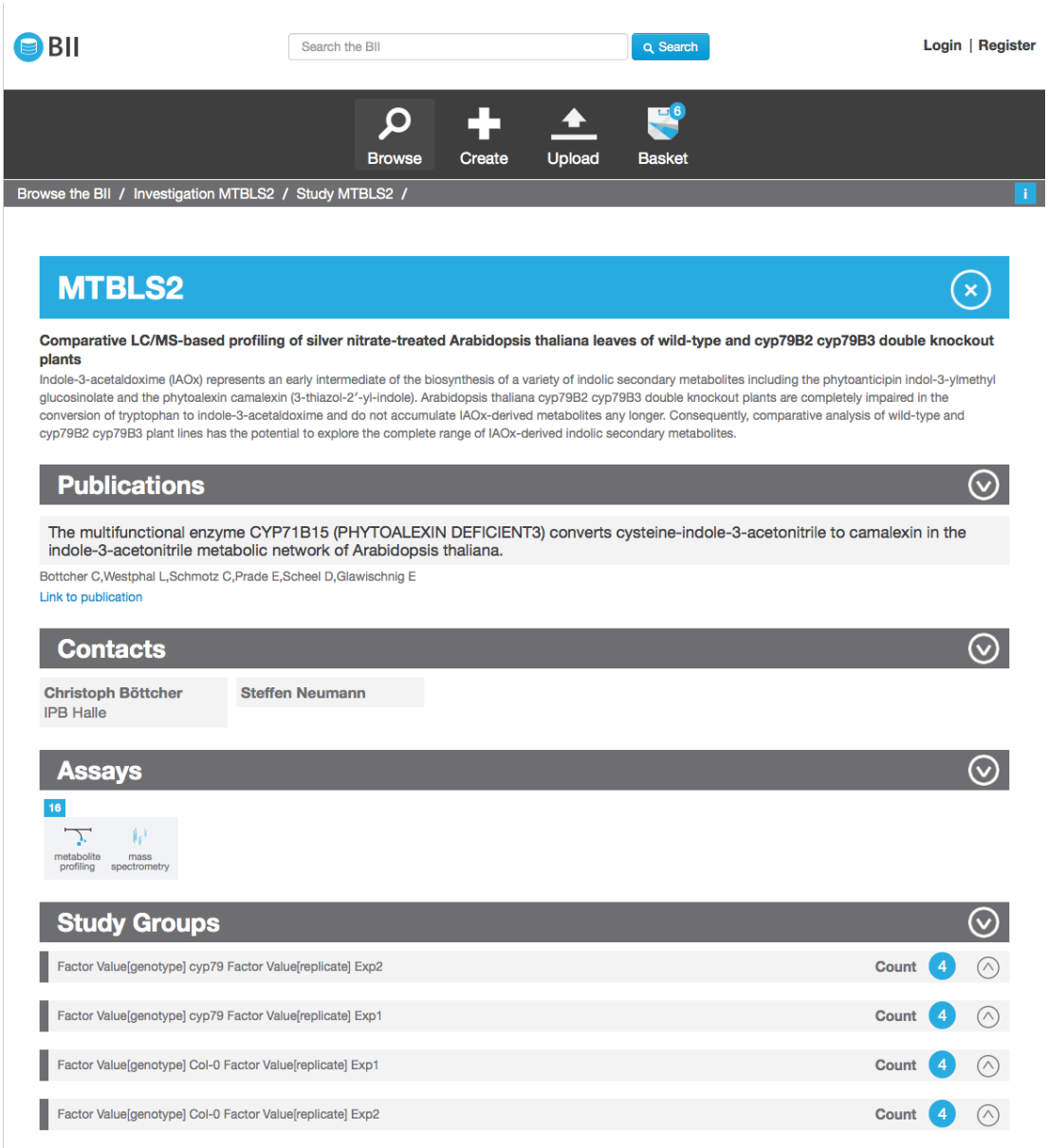


Figure 3: View of the MTBLS2 study in Bio-GraphIIne, a Django-powered web application with a graph database backend storing RDF named graphs generated by LinkedISA software component.

3.2.4 Development of an RDF-ified MassBank SPARQL endpoint

We applied a subset of the guidelines to our efforts at IPB to create a semantic web triple store, making Mass Bank core data available in a LOD fashion. To structure the data in such a triple store, we had to develop a Subject-Predicate-Object (SPO) triple style RDF model. Ontology is used to further formalize the generated RDF model.

The following graphic shows the current RDF MassBank schema as a graph of SPO RDF triples (Fig. 5):

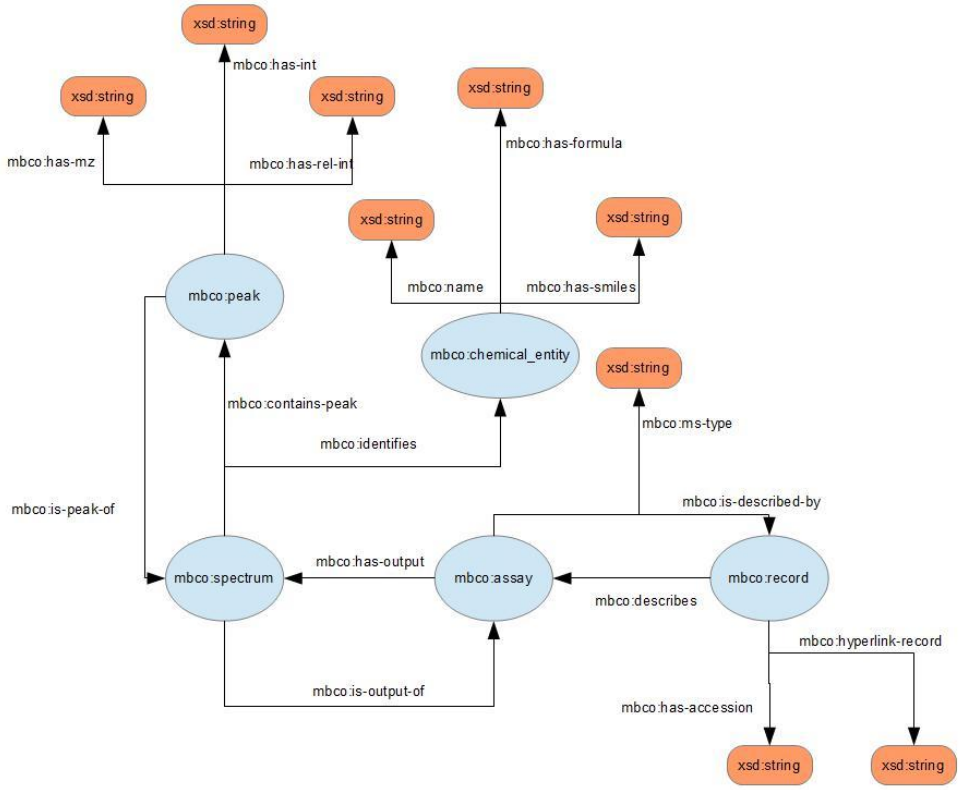


Figure 4: Current simple and intuitive RDF graph describing the MassBank LOD schema.

This model currently consists of 5 classes and 17 predicates, and is defined in the Terse Triple Language, the turtle RDF syntax. The following listing is an excerpt:

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix mbco: <http://www.ipb-halle.de/ontology/mbco#> .

mbco:record a rdfs:Class ;
   rdfs:isDefinedBy <http://www.ipb-halle.de/ontology/mbco#record> ;
   rdfs:label "Record" ;
   rdfs:comment "A class for massbank records" ;
   rdfs:subClassOf <http://semanticscience.org/resource/SIO_000088> .

mbco:assay a rdfs:Class ;
   rdfs:isDefinedBy <http://www.ipb-halle.de/ontology/mbco#assay> ;
   rdfs:label "Assay" ;
   rdfs:comment "The assay describing the mass spectrometry experiment
that is described in the record" .

mbco:spectrum a rdfs:Class ;
   rdfs:isDefinedBy <http://www.ipb-halle.de/ontology/mbco#spectrum> ;
   rdfs:label "Spectrum" ;
   rdfs:comment "Information of the spectrum generated by an assay" .
```

3.2.5 Prototype SPARQL endpoints

Two SPARQL endpoint prototypes are already in use, one at the University of Oxford e-Research Centre and another one at IPB.

3.2.5.1 Oxford MetaboLights Endpoint

The endpoint relies on Virtuoso stack and Sesame was also evaluated. It is used to host the converted content of the EMBL-EBI MetaboLights repository and test representation options, Sparql queries and query optimization.

RDF Triple Store Faceted Browser:
http://newt.oerc.ox.ac.uk:8890/fct/

Triple Store SPARQL Endpoint:
http://newt.oerc.ox.ac.uk:8890/sparql

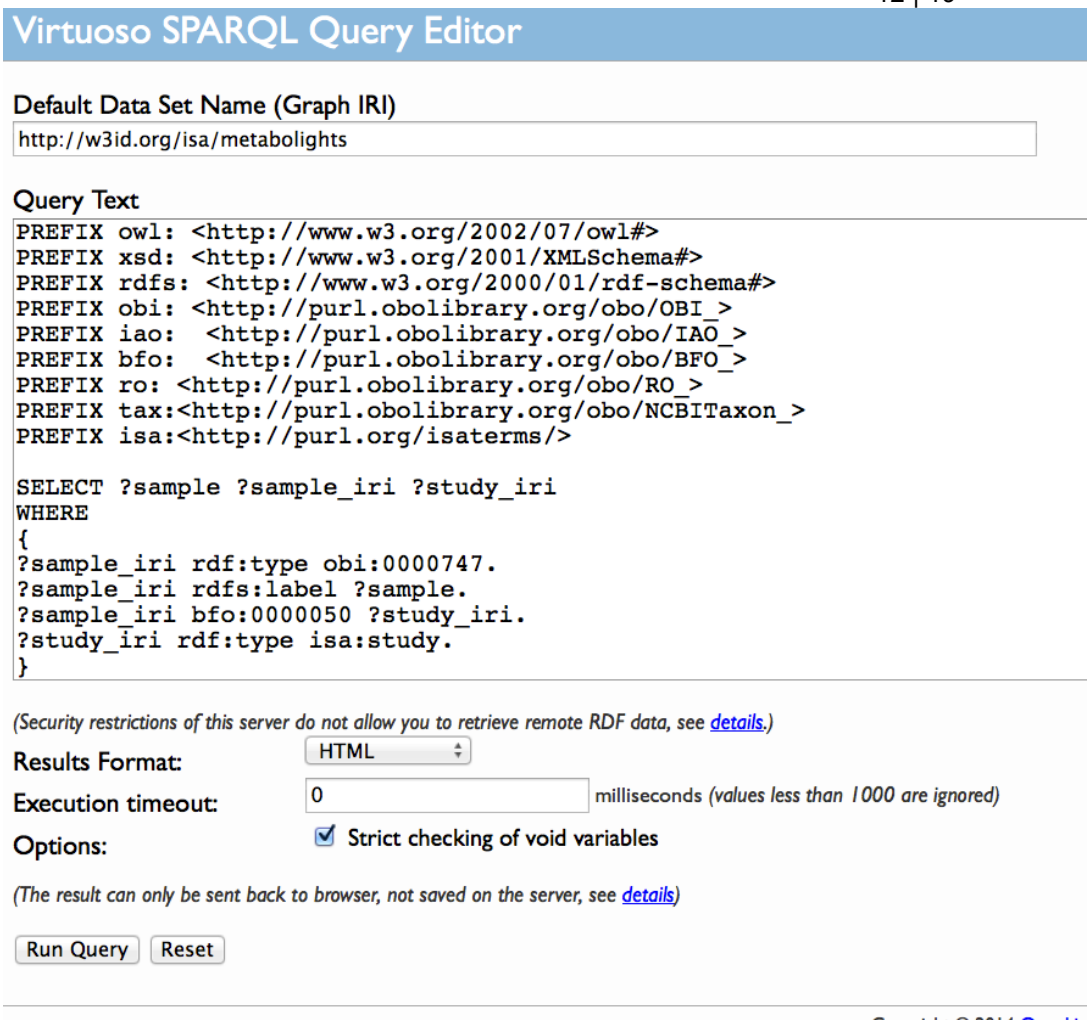


Figure 5: An example SPARQL query for ISA MetaboLights studies

3.2.5.1 IPB MassBank Endpoint

The IPB has installed the Open Source edition of the Virtuoso triple store and SPARQL endpoint. In addition to the triple store and SPARQL query interface, we implemented several prototypes to test automatic data conversion and queries (Fig 6).

We imported the MassBank, ChEBI and MetaboLights (created via linkedISA) data into the triple store, and created several example queries.

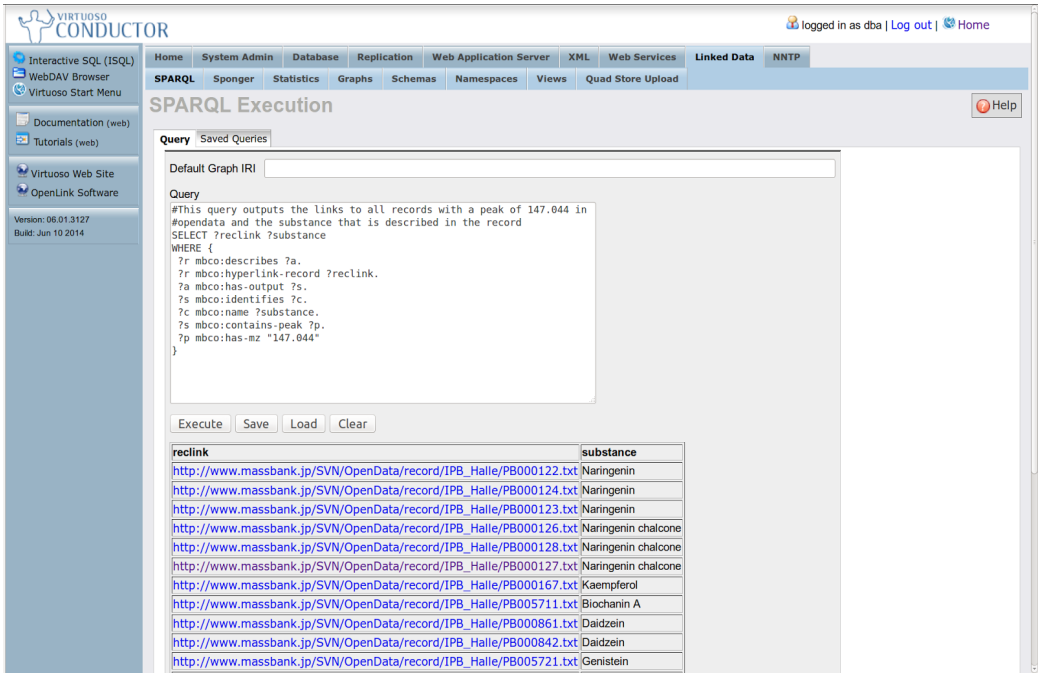


Figure 6: IPB MassBank triple store containing the converted MassBank data. The screenshot shows an example query for reference spectra containing a peak with an m/z of 147.644 and the resulting records.

3.2.6 IPB Access to resources

All source files for the IPB endpoints and the RDF guideline are available on the project Github pages, together with an accompanying readme file.

GitHub:

https://github.com/sneumann/SemanticMetabolomics

Guideline document:

http://www.cosmos-fp7.eu/system/files/presentation/RDFicationguidelineforMetabolomicsLinkedDatacreation.pdf

The production-mode SparQL endpoints will be made public later during the COSMOS project.

3.3 Next steps

On the terminological side, we will further need to develop the "inventory" of terms required to express knowledge about metabolomics experiments, their processing and results.

The major next step will be exemplary queries across several geographically distributed endpoints, to showcase the benefits of Linked Open Data. A Semantic Metabolomics Workshop is planned for 2015.

Further efforts will be devoted to engage more partners and data providers to investigate RDF and SPARQL as future additions to their services, and reconcile semantic framework used. In addition, we need to bring Metabolomics to the attention of the existing LOD community. To that end, Daniel Schober and Michael van Vliet plan to attend the "Semantic web applications and tools for Life Science" (http://www.swat4ls.org/) workshop in December 2014 in Berlin.

4 Publications

- *Bio-GraphIIn: a graph-based, integrative and semantically-enabled repository for life science experimental data.* Alejandra Gonzalez-Beltran , Eamonn Maguire, Pavlos Georgiou, Susanna-Assunta Sansone, Philippe Rocca-Serra. Proceedings of *NETTAB 2013, EMBNet Journal 2013. DOI:* http://dx.doi.org/10.14806/ej.19.B.728
- *linkedISA: semantic representation of ISA-Tab experimental metadata.* Alejandra Gonzalez-Beltran, Eamonn Maguire, Susanna-Assunta Sansone, Philippe Rocca- Serra. BMC Bioinformatics 2014, *in press.*

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
8. EMBL-EBI	1		
11. IPB	2		
4. IMPERIAL	1		
14:UOXF			
8:MPG	2		
2:LU	1		
	7	2	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