

## 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 Program; [INFRA-2011-2.3.2.] Implementation of common solutions for a cluster of ESFRI infrastructures in the field of "Life sciences"
Deliverable title:	D2.8 Guideline Document on RDF and SPARQL for metabolomics resources
WP No.	2
Lead Beneficiary:	11. IPB
WP Title	Standards Development

Contractual delivery date:	30 September 2014
Actual delivery date:	30 October 2014
WP leader:	Steffen Neumann (Daniel Schober) 11. IPB
Contributing partner(s):	11. IPB, 1.EMBL-EBI, 3. MRC, 2. MPG, 14 UOXF

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

## Content

- [1 Executive summary](#)
- [2 Project objectives](#)
- [3 Detailed report on the deliverable](#)
  - [3.1 Background](#)
  - [3.2 Description of Work](#)
    - [3.2.1 Description of Use Cases in Metabolomics](#)
    - [3.2.2 Competency Questions for the Metabolomics Use Case](#)
    - [3.2.3 Conversion of Metabolights Metadata description to RDF using LinkedISA component](#)
    - [3.2.4 Development of an RDF-ified MassBank SPARQL endpoint](#)
    - [3.2.5 Prototype SPARQL endpoints](#)
      - [3.2.5.1 Oxford Metabolights Endpoint](#)
    - [3.2.6 Access to resources](#)
  - [3.3 Next steps](#)
- [4 Publications](#)
- [5 Delivery and schedule](#)
- [6 Adjustments made](#)
- [7 Efforts for this deliverable](#)
- [Background information](#)
- [References](#)

## 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 need 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 contributed the following objectives:

No.	Objective	Yes	No
4	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.	X	

## 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.

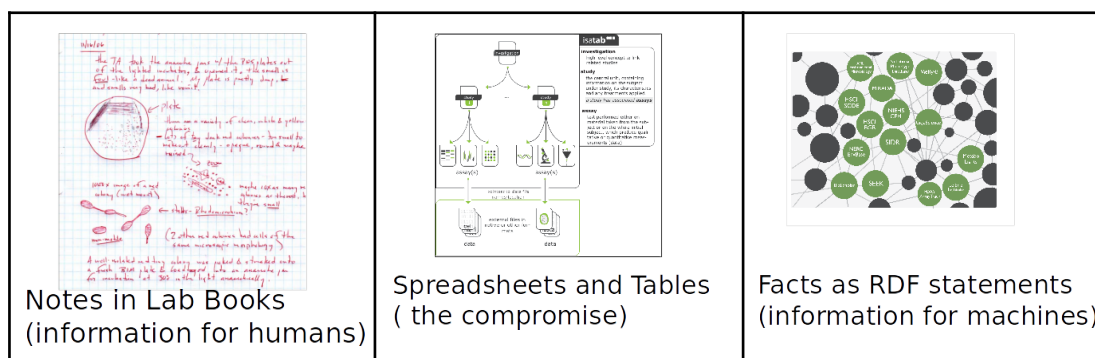
To demonstrate the feasibility, we create exemplary semantic web query endpoints and will later connect these for distributed integrative querying. The EBI, MPG and IPB will augment their MetaboLights, GMD and MassBank databases with LOD resources.

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 linkability 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 EBI's RDF guideline<sup>1</sup> and the Bio2RDF guidelines<sup>2</sup>. We also reviewed guideline documents by general policy providers like the W3C consortium.



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

<sup>1</sup> <http://www.ebi.ac.uk/rdf/rdf-first-principles>

<sup>2</sup> <https://github.com/bio2rdf/bio2rdf-scripts/wiki/Bio2RDF-Release-2-ICBO-Tutorials>

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 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 Brassicaceae 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<sup>3</sup>, and the bio2rdf project<sup>4</sup> at Carlton University, CA.

### 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 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

<sup>3</sup> <http://www.ebi.ac.uk/about/news/press-releases/RDF-platform>

<sup>4</sup> <http://bio2rdf.org/>

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.

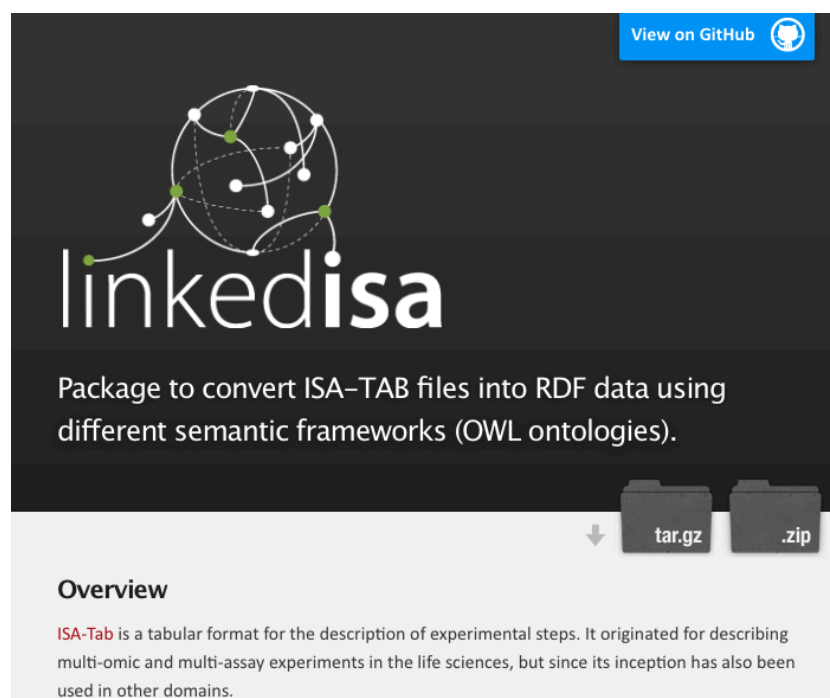
The LinkedISA conversion component produced by UOXF allows the transformation of an ISA formatted study into an RDF named graph, and offers the ability to carry out further validation checks and automatically augment annotation without user intervention, by taking advantage of the underlying semantic 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 Metabolomic 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/>



[Login](#) | [Register](#)

 Browse
  Create
  Upload
  Basket

[Browse the BII](#) / [Investigation MTBLS2](#) / [Study MTBLS2](#) /

## MTBLS2

**Comparative LC/MS-based profiling of silver nitrate-treated Arabidopsis thaliana leaves of wild-type and cyp79B2 cyp79B3 double knockout plants**

Indole-3-acetaldoxime (IAOx) represents an early intermediate of the biosynthesis of a variety of indolic secondary metabolites including the phytoalexin indol-3-ylmethyl glucosinolate and the phytoalexin camalexin (3-thiazol-2'-yl-indole). Arabidopsis thaliana cyp79B2 cyp79B3 double knockout plants are completely impaired in the conversion of tryptophan to indole-3-acetaldoxime and do not accumulate IAOx-derived metabolites any longer. Consequently, comparative analysis of wild-type and cyp79B2 cyp79B3 plant lines has the potential to explore the complete range of IAOx-derived indolic secondary metabolites.

### Publications

The multifunctional enzyme CYP71B15 (PHYTOALEXIN DEFICIENT3) converts cysteine-indole-3-acetonitrile to camalexin in the indole-3-acetonitrile metabolic network of Arabidopsis thaliana.

Böttcher C, Westphal L, Schmotz C, Prade E, Scheel D, Glawischnig E

[Link to publication](#)



### Contacts

[Christoph Böttcher](#)  
 IPB Halle

[Steffen Neumann](#)

### Assays

16

 metabolite profiling
  mass spectrometry

### Study Groups

Factor Value[genotype] cyp79 Factor Value[replicate] Exp2	Count	4	⬆
Factor Value[genotype] cyp79 Factor Value[replicate] Exp1	Count	4	⬆
Factor Value[genotype] Col-0 Factor Value[replicate] Exp1	Count	4	⬆
Factor Value[genotype] Col-0 Factor Value[replicate] Exp2	Count	4	⬆

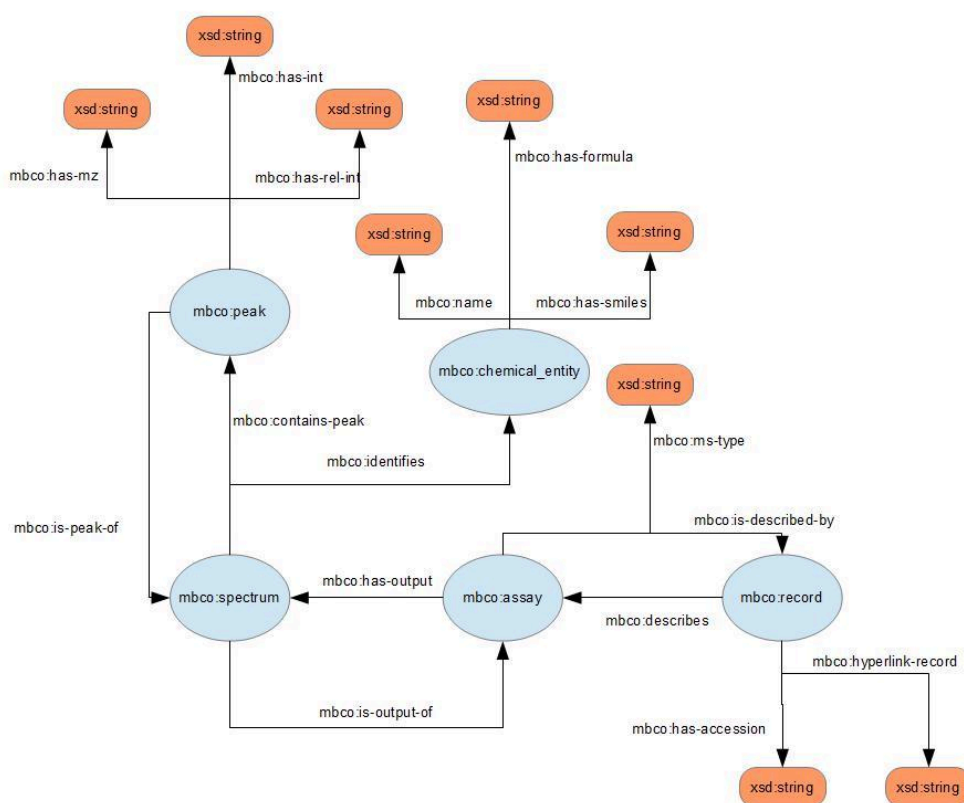
**Figure 3:** View of the MTBLS2 study in Bio-GraphLine, a Django-powered web application with a graph database backend storing RDF named graphs generated by LinkedISA software component.

### 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. An 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.

**Virtuoso SPARQL Query Editor**

**Default Data Set Name (Graph IRI)**

http://w3id.org/isa/metabolights

**Query Text**

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX obi: <http://purl.obolibrary.org/obo/OBI_>
PREFIX iao: <http://purl.obolibrary.org/obo/IAO_>
PREFIX bfo: <http://purl.obolibrary.org/obo/BFO_>
PREFIX ro: <http://purl.obolibrary.org/obo/RO_>
PREFIX tax:<http://purl.obolibrary.org/obo/NCBITaxon_>
PREFIX isa:<http://purl.org/isaterms/>

SELECT ?sample ?sample_iri ?study_iri
WHERE
{
  ?sample_iri rdf:type obi:0000747.
  ?sample_iri rdfs:label ?sample.
  ?sample_iri bfo:0000050 ?study_iri.
  ?study_iri rdf:type isa:study.
}
```

(Security restrictions of this server do not allow you to retrieve remote RDF data, see [details](#).)

**Results Format:** HTML

**Execution timeout:** 0 milliseconds (values less than 1000 are ignored)

**Options:** ☒ Strict checking of void variables

(The result can only be sent back to browser, not saved on the server, see [details](#))

Run Query
Reset

**Figure 5:** An example SPARQL query for ISA Metabolights studies

#### RDF Triple Store Faceted Browser:

<http://newt.oerc.ox.ac.uk:8890/fct/>

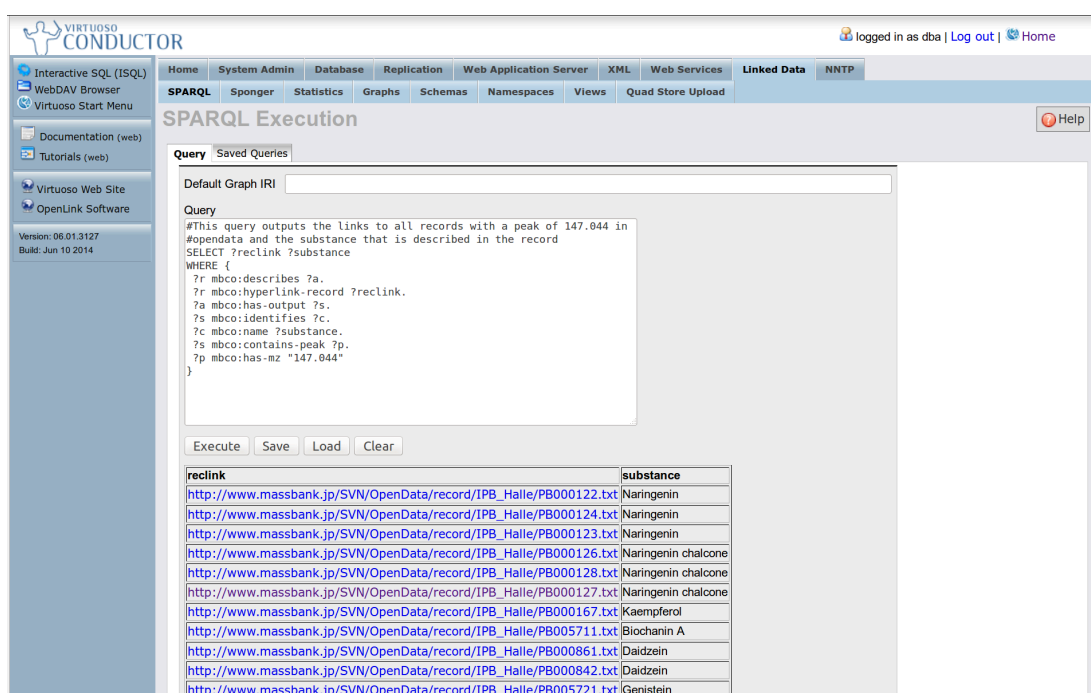
#### Triple Store SPARQL Endpoint:

<http://newt.oerc.ox.ac.uk:8890/sparql>

### 3.2.5.2 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.



The screenshot shows the Virtuoso Conductor SPARQL Execution interface. The query is as follows:

```
#This query outputs the links to all records with a peak of 147.044 in
#opendata and the substance that is described in the record
SELECT ?relink ?substance
WHERE {
  ?r mbco:describes ?a.
  ?r mbco:hyperlink-record ?relink.
  ?a mbco:has-output ?s.
  ?s mbco:identifies ?c.
  ?c mbco:name ?substance.
  ?s mbco:contains-peak ?p.
  ?p mbco:has-mz "147.044"
}
```

The results table shows the following records:

relink	substance
<a href="http://www.massbank.jp/SVN/OpenData/record/IPB_Halle/PB000122.txt">http://www.massbank.jp/SVN/OpenData/record/IPB_Halle/PB000122.txt</a>	Naringenin
<a href="http://www.massbank.jp/SVN/OpenData/record/IPB_Halle/PB000124.txt">http://www.massbank.jp/SVN/OpenData/record/IPB_Halle/PB000124.txt</a>	Naringenin
<a href="http://www.massbank.jp/SVN/OpenData/record/IPB_Halle/PB000123.txt">http://www.massbank.jp/SVN/OpenData/record/IPB_Halle/PB000123.txt</a>	Naringenin
<a href="http://www.massbank.jp/SVN/OpenData/record/IPB_Halle/PB000126.txt">http://www.massbank.jp/SVN/OpenData/record/IPB_Halle/PB000126.txt</a>	Naringenin chalcone
<a href="http://www.massbank.jp/SVN/OpenData/record/IPB_Halle/PB000128.txt">http://www.massbank.jp/SVN/OpenData/record/IPB_Halle/PB000128.txt</a>	Naringenin chalcone
<a href="http://www.massbank.jp/SVN/OpenData/record/IPB_Halle/PB000127.txt">http://www.massbank.jp/SVN/OpenData/record/IPB_Halle/PB000127.txt</a>	Naringenin chalcone
<a href="http://www.massbank.jp/SVN/OpenData/record/IPB_Halle/PB000167.txt">http://www.massbank.jp/SVN/OpenData/record/IPB_Halle/PB000167.txt</a>	Kaempferol
<a href="http://www.massbank.jp/SVN/OpenData/record/IPB_Halle/PB005711.txt">http://www.massbank.jp/SVN/OpenData/record/IPB_Halle/PB005711.txt</a>	Biochanin A
<a href="http://www.massbank.jp/SVN/OpenData/record/IPB_Halle/PB000861.txt">http://www.massbank.jp/SVN/OpenData/record/IPB_Halle/PB000861.txt</a>	Daidzein
<a href="http://www.massbank.jp/SVN/OpenData/record/IPB_Halle/PB000842.txt">http://www.massbank.jp/SVN/OpenData/record/IPB_Halle/PB000842.txt</a>	Daidzein
<a href="http://www.massbank.jp/SVN/OpenData/record/IPB_Halle/PB005721.txt">http://www.massbank.jp/SVN/OpenData/record/IPB_Halle/PB005721.txt</a>	Genistein

**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 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-GraphIn: 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

None.

## 7 Efforts for this deliverable

Institute	Person-months (PM)		Period
	actual	estimated	
11. IPB	1.2		

1. EMBL-EBI	0.3		
14. UOXF	0.5		
Total	2	2	

## Background information

### UPDATE WITH WP INFO

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-Halle  
Participants: Michael Wilson, Wishart Group, Edmonton Canada, 1.EBI , 14 UOXF, 12 UB2, 13 UBHam

In this deliverable D 2.4 we have coordinated efforts from multiple international groups who are working in NMR and metabolomics related software to design and establish a vendor agnostic nmrML data format. The standards development work package (COSMOS WP2) here delivers the essential exchange standard for NMR-based metabolomics raw data.

<b>Work package number</b>	WP2	<b>Start date or starting event:</b>	November 2012
----------------------------	-----	--------------------------------------	---------------

<b>Work package title</b>	WP2, Standards Development
<b>Activity Type</b>	Coordination, prototype

<b>Participant number</b>	No: partner	11. IPB	No: partner	No: partner	No: partner	No: partner	No: partner	No: partner
<b>Person months per participant</b>	XX		XX	XX	XX	XX	XX	XX

<b>Objectives</b> Insert objective 1 Insert objective 2
<b>Description of work and role of participants</b> Insert WP description, tasks etc.
<b>Deliverables</b>

No.	Name	Due month
DX.X	Insert deliverable title	X
DX.X	Insert deliverable title	X

## References