

Preliminary Guidelines Towards Semantic Metabolomics Resources

exemplified and explained according
to the MassBank LOD prototype platform at IPB Halle (D. Schober, S. Neumann)
to the Metabolights LOD conversion by LinkedISA at Oxford (A. Gonzalez-Beltran, P. Rocca-Serra)

Disclaimer:

This document is a draft and work in progress. We capture here the experience gathered by two subprojects, each looking at specialized resources storing Metabolomics data.

The reader will therefore find a set of outlines and recommendations stemming from this work and documentation on the key steps and practices in order to produce optimal RDF linked data to facilitate resource querying, interoperability and data integration.

*Its content will move it onto the COSMOS Semantic Metabolomics wiki at
<https://github.com/sneumann/SemanticMetabolomics/wiki>*

Contents

[Preliminary Guidelines](#)

[Towards Semantic Metabolomics Resources](#)

[Contents](#)

[Scope of this document](#)

[Overall set up](#)

[Use cases](#)

[CQs \(Competency Questions\)](#)

[Selection of Resource Modules \(Data providers\) for which synergism is expected](#)

[Core module example LOD graphs](#)

[MassBank Resource Module:](#)

[The mbco.owl: mass bank core ontology](#)

[Metabolights Resource Module](#)

[Requirement analysis \(with established best practices followed\)](#)

[Virtuoso and SPARQL endpoint setup](#)

[Example Sparql queries & output](#)

[Abstracted \(general\) workflow for creating linked Data resources](#)

[Best Practice Documents for Linked Data that we adhere to](#)

[Literature references](#)

[SemWeb & LOD Tutorials](#)

[Useful hyperlinks](#)

[Tools](#)

[General](#)

[RDFizers](#)
[Semantic Web Search Engines](#)
[Open Linked Data Browsers](#)
[Validators](#)

Scope of this document

This document captures the preliminary experience of producing Resource Description Framework (RDF) graphs compatible with a linked open data (LOD) approach. To this end, our two groups (OeRC and IPB) have worked on converting two Metabolomics resources, namely Metabolights and MassBank respectively, into prototypic RDF resources mirroring essential data from each of these resources along the Semantic Web and LOD paradigm. We chose those two resources as representing the main types of data sought to be represented, namely experimental metadata from ISA-Tab files served by Metabolights and identified metabolites and their spectral signatures as served by Massbank.

The experience gained can later serve as a test case for the wider COSMOS efforts towards a semantic web of metabolomics data resources. The production version hence later needs to be reimplemented in a more decentralized way accessing remote endpoints and will be expanded e.g. to fulfil requirements of the NMR world as well, i.e. by linking to HMDB data.

This document also intends to collect established RDF best practices and apply them for our metabolomics use case. Adherence to these established guidelines will ease peer-compliance and later integration with existing LOD efforts. Here, in particular the W3C HCLS, OBO Foundry, Bio2RDF community and the EBI RDF resources provided guidance. Additional Semantic Web best practices were followed along the sources provided below.

Overall set up

For each resource, we followed slightly different approaches:

Metabolights:

ISA-Tab formatted experiments served by EMBL-EBI have been converted to RDF named graphs using LinkedISA software developed at OeRC. A Virtuoso RDF triple store collected the data and makes it web-accessible.

MassBank:

We here describe an experimental RDF dump of Massbank core data that was generated in a centralized local approach via makefiles that autoconvert selected MassBank parts into one large RDF triple store, again published by Virtuoso for SPARQL querying.

Use cases

Our main use case is to deliver SPARQL endpoints for metabolomics core data like MassBank spectral reference data and metabolic assays, which allows to be integrated with existing

endpoints in related domains, e.g. endpoints on small molecules like ChEBI, or an endpoint for describing plant or animal species. The basic assumption here is that scientific data providers should deliver their data to the community in an open and integratable fashion, namely the W3C recommended RDF syntax, to facilitate data re-use and ease contextual enrichment.

CQs (Competency Questions)

We have collected a set of competency questions from COSMOS stakeholders and putative users, to drive and inform our development. Those competency questions encompass a range of granularity levels, from top level descriptions of experimental setup down to important domain specific descriptors for molecular identification. These CQs serve later evaluation on conceptual completeness & relational coverage of the model:

Competency questions:

1. Select all MassBank records with a certain peak
2. Select all MassBank records about certain ChEBI compounds
3. Select all Metabolights studies which mention a metabolite for which there is a MassBank record
4. Select all compounds from Metabolights, which are mentioned for Brassicaceae species
5. Select all MassBank records of samples measured in Germany. This will require annotation of the records with gazetteer ontology terms (<http://bioportal.bioontology.org/ontologies/GAZ/>), and a selection of everything that is "located_in" e.g. Germany (<http://www.ebi.ac.uk/ontology-lookup/browse.do?ontName=GAZ&termId=GAZ:00002646&termName=Germany>). This is similar to the NCBI taxonomy query for all brassicaceae.

Future questions:

1. Select all compounds from Metabolights, with has_role insecticide (CHEBI:24852)
2. Select all MassBank records for Molecules that interact with Protein/Enzyme X (using e.g. <http://stitch.embl.de/>)? Using the Stitch sparql endpoint would allow to connect metabolomics with Proteomics.
3. Give me all metabolites intensities measured from the same samples, but in different assays (positive/negative mode, MS/NMR, using mzMine/XCMS, ...)

Selection of Resource Modules (Data providers) for which synergism is expected

Resources we intend to interlink for contextual enrichment at a later stage are, but are not limited to:

1. MassBank for MS reference spectra: Own local dump
2. Metabolights for metabolism related assays and investigations
3. ChEBI for small molecule descriptions: Accessible via NCBO SPARQL endpoint(on <http://sparql.bioontology.org/> it is <http://bioportal.bioontology.org/ontologies/47013>) or Bio2RDF ?
4. NCI Tax or NEWT for a species taxonomy

5. HMDB for NMR reference spectra

Further potentially interesting resources are available for Tissues, SampleTypes and Statistical models.

Core module example LOD graphs

MassBank Resource Module:

A typical MassBank Record instance:

Here we exemplarily show a native MassBank source entry. We have underlined the core elements that our schema currently captures and hence can be queried for:

ACCESSION: UF005001

RECORD_TITLE: Diphenylamine; ESI-ITFT; direct injection; MS2; CE: 25; CID; R=60000; [M+H]⁺

DATE: 2013.11.03

AUTHORS: Krauss M, Schulze T, Department of Effect-Directed Analysis, Helmholtz Centre for Environmental Research - UFZ GmbH, Leipzig, Germany

LICENSE: CC BY-SA

COPYRIGHT: Copyright (C) 2012, 2013 UFZ, Leipzig, Germany

COMMENT: CONFIDENCE standard compound

COMMENT: INTERNAL_ID 50

CH\$NAME: Diphenylamine

CH\$NAME: N-Phenylaniline

CH\$COMPOUND_CLASS: N/A; Environmental Standard

CH\$FORMULA: C12H11N

CH\$EXACT_MASS: 169.08915

CH\$SMILES: N(C1=CC=CC=C1)C1=CC=CC=C1

CH\$IUPAC: InChI=1S/C12H11N/c1-3-7-11(8-4-1)13-12-9-5-2-6-10-12/h1-10,13H

CH\$LINK: CAS 122-39-4

CH\$LINK: KEGG C11016

CH\$LINK: PUBCHEM CID:11487

CH\$LINK: INCHIKEY DMBHRLKUKUOEG-UHFFFAOYSA-N

CH\$LINK: CHEMSPIDER 11003

AC\$INSTRUMENT: LTQ Orbitrap XL Thermo Scientific

AC\$INSTRUMENT_TYPE: ESI-ITFT

AC\$MASS_SPECTROMETRY: MS_TYPE MS2

AC\$MASS_SPECTROMETRY: IONIZATION ESI

AC\$MASS_SPECTROMETRY: ION_MODE POSITIVE

AC\$MASS_SPECTROMETRY: FRAGMENTATION_MODE CID

AC\$MASS_SPECTROMETRY: COLLISION_ENERGY 25

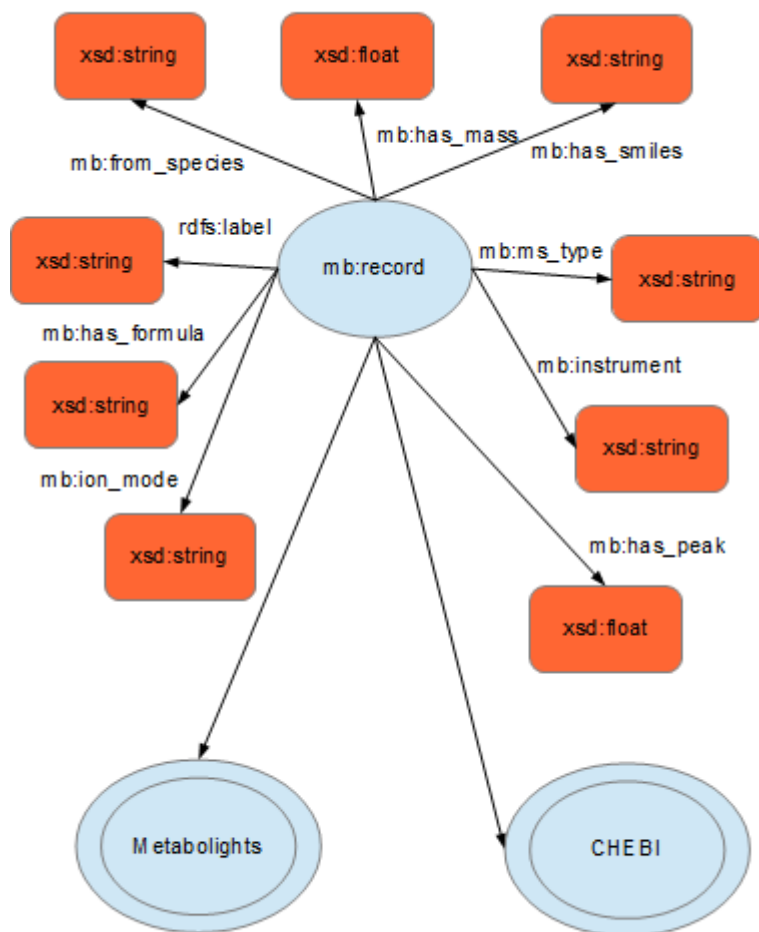
AC\$MASS_SPECTROMETRY: RESOLUTION 60000

AC\$CHROMATOGRAPHY: COLUMN_NAME

AC\$CHROMATOGRAPHY: FLOW_GRADIENT 90/10 at 0 min, 80/20 at 0.7 min, 100/0 at 5.0

min, 90/10 at 10 min, 90/10 at 11 min
AC\$CHROMATOGRAPHY: FLOW_RATE 200 uL/min
AC\$CHROMATOGRAPHY: RETENTION_TIME 4.8 min
AC\$CHROMATOGRAPHY: SOLVENT A water with 0.1% formic acid
AC\$CHROMATOGRAPHY: SOLVENT B methanol with 0.1% formic acid
MS\$FOCUSED_ION: BASE_PEAK 149.0025
MS\$FOCUSED_ION: PRECURSOR_M/Z 170.0964
MS\$FOCUSED_ION: PRECURSOR_TYPE [M+H]⁺
MS\$DATA_PROCESSING:
MS\$DATA_PROCESSING: REANALYZE Peaks with additional N2/O included
MS\$DATA_PROCESSING: WHOLE RMassBank 1.0.0 & 1.5.0
PK\$ANNOTATION: m/z tentative_formula formula_count mass error(ppm)
92.0491 C6H6N⁺ 1 92.0495 -4.08
PK\$NUM_PEAK: 1
PK\$PEAK: m/z int. rel.int.
92.0491 1608 999
//

Based on this core data our students came up with an initial naive graph capturing the core information along one single RDF node. Note that "rdfs:label" points to the accession number. Chebi and Metabolights are marked as super-entities, as the exact relation is yet still unknown. The current predicates in the graph are most likely just placeholders and we'll replace many of them with already existing predicates, e.g. from SIO or RO.



This initial MassBank RDF schema turned out to be oversimplified and needed refinement along the following semantic web best practices:

- Strive for 'Things, not Strings' philosophy:
<http://googleblog.blogspot.co.uk/2012/05/introducing-knowledge-graph-things-not.html>
 Try to make as much URIs out of Strings as you can, i.e. for Species use NCITax controlled vocabulary instead of ambiguous freetext. Make the used entities de-referenceable by HTTP URIs that show the classes from the used vocabularies/ontologies.
- Implement a mass bank (mb) core ontology (mbco.owl) for semantic specification. This can be an extension of the SIO ontology (<https://code.google.com/p/semanticscience/wiki/SIO>), which is widely used in Bio2RDF (<http://bio2rdf.org>), chembl core ontology and e.g. SADI (<http://sadiframework.org>). An mbco draft is found in the Git/ontology folder.
- Refactor the mass bank model into at least three ontologically justified basic classes to get semantically more clean and avoid semantic overloading. Sloppy assertions such as

“record *has_mass* float” (an invalid proxy-shortcut) are bound to confuse humans and computers in particular, as it is odd to say that an information content entity like a DTB record has a mass. Such blur in meaning due to conflated ontological concepts must be avoided.

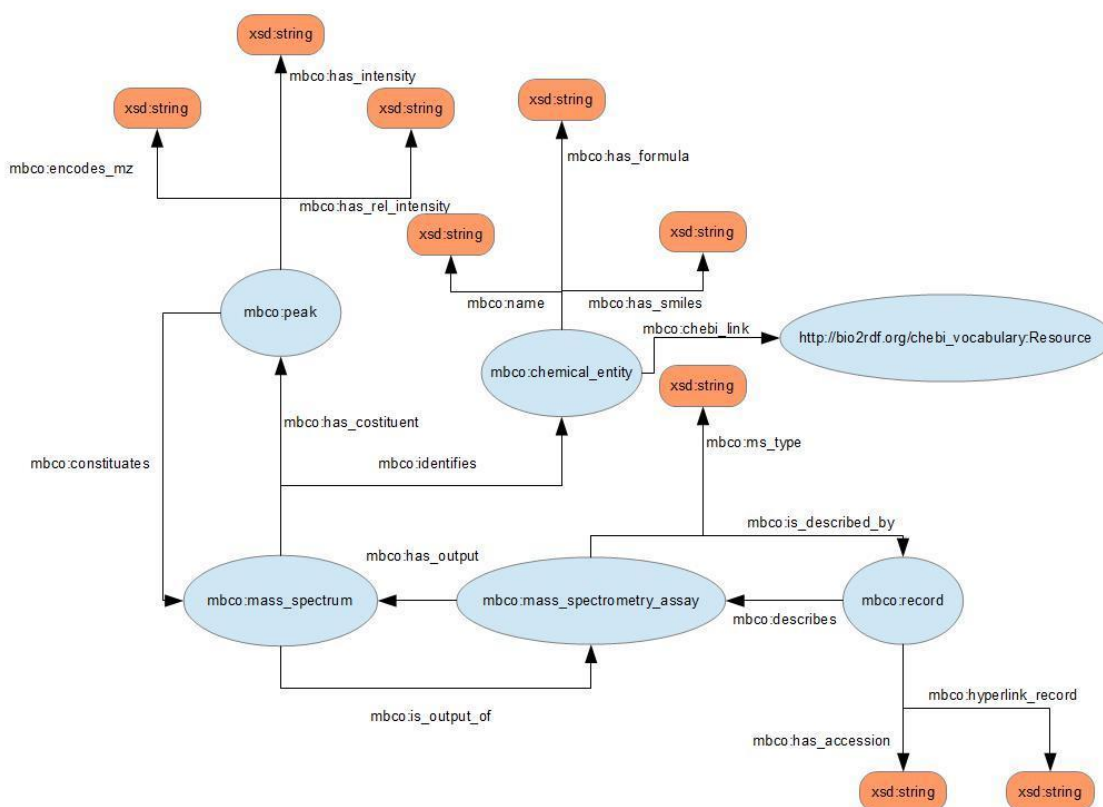
Based on these guidelines we created a first sketch of a node and property lists for the MassBank RDF schema:

- `sio:experiment/mb:Chemical characterisation assay/mb:mass spectrometry assay`
- `iao:information content entity/mb:mass bank record`
- `bfo:object/mb:chemical entity`
- `iao:information content entity/mb:peak`

Only then can we start to assign ontologically clean and rigid triples, e.g.

- `mb:mass spectrometry assay has_output mass bank record`
- `mass bank record describes peak`
- `peak identifies chemical entity`
- `chemical entity has_smiles_code smiles code`
- **`chemical entity has_mass float (the correct form of the faulty assertion above)`**

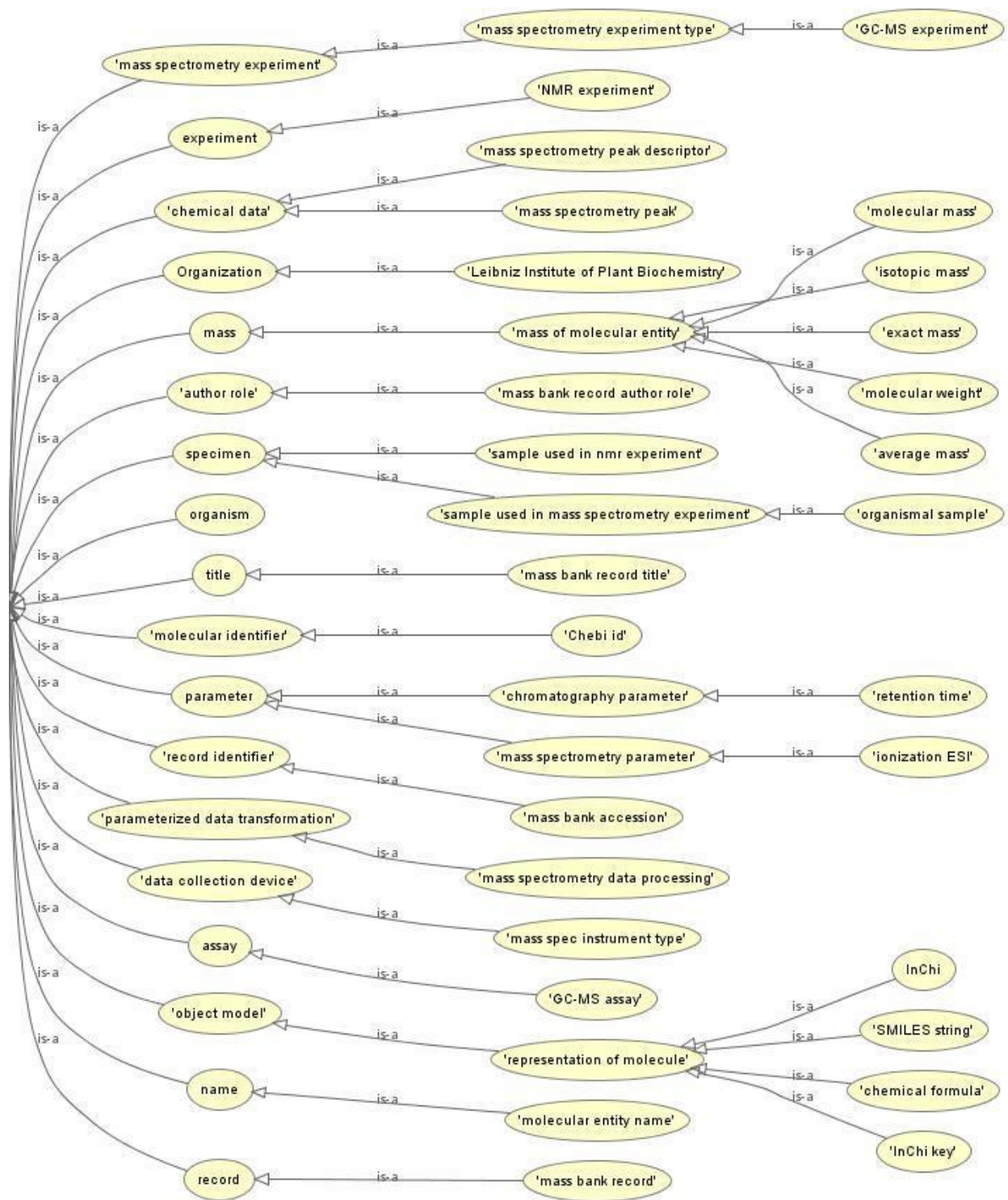
We implemented the above in the current updated version of MassBank core RDF schema:



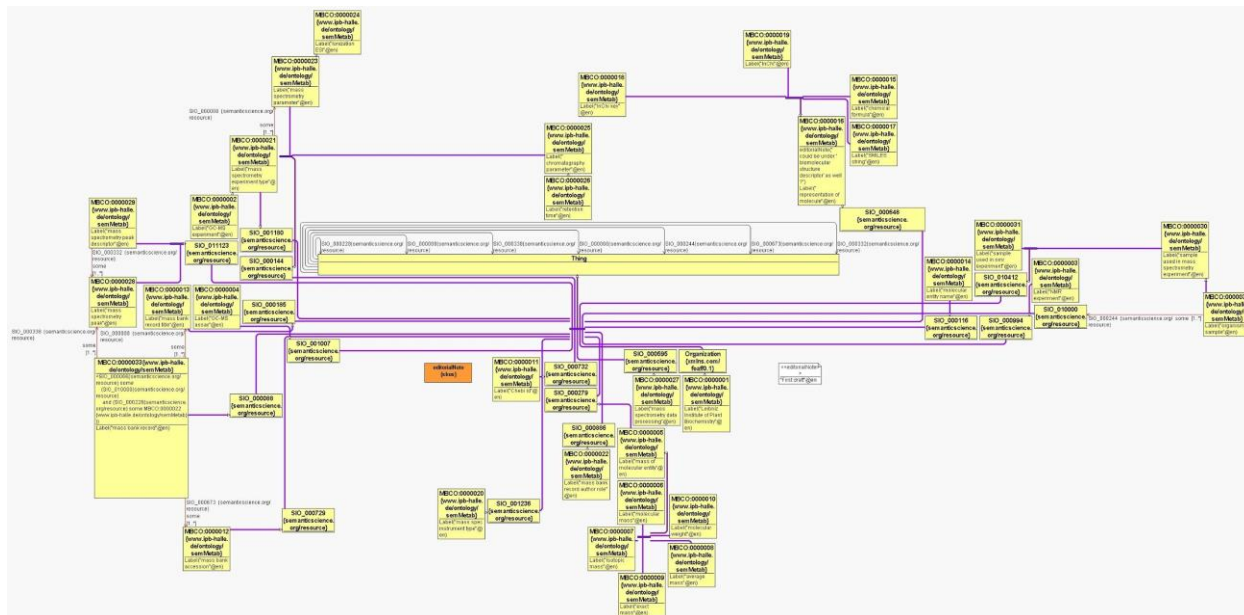
This model currently consists of 5 classes and 17 predicates and is described in our Git at [SemanticMetabolomics\definitions\mb.ttl](#) (Terse Triple Language, the turtle RDF syntax). It already allows the answering of the MassBank related CQs (e.g. CQ1, as shown in the example below).

The mbco.owl: mass bank core ontology

We generated an owl file which imports the SIO ontology (used by bio2RDF) as semantically traceable upper level model. The mbco ontology can be found in the git ontology folder. Its main taxonomy reads as follows:



Its relational structure can be found in this OWL Gred graph (a High Res Figure can be found in the Git/Docs folder):



Metabolights Resource Module

The ISA-Tab configured Metabolights entries were RDF-converted using the OeRC proprietary LinkedISA converter, exploiting a mapping file using OBO Foundry resources, namely OBI for experimental metadata, CHEBI for chemical entities, GO for molecular function, process and cellular localization entities. These terminological resources all belong to the foundry and are widely used by the community. Of importance, LinkedISA provides the possibility to substitute the OBO Foundry based mappings with SIO based mappings and supports Prov-O, an RDF model for provenance tracking.

Requirement analysis (with established best practices followed)

We here list some of the guiding considerations on a per Issue basis:

- Why these NS/resource modules ? → Domain constrained, robustness, resource persistence/funding, usage, ...
- Why were these core/hub resource nodes chosen ? → CQ justification, robustness, availability, will always be a pragmatic compromise. Need to find cutoff between complexity & performance and simplicity & ability to create SPARQL queries...
- Why expressivity is RDF (avoid rdf containers and blank nodes), less Owl-->SPARQL optimized expressivity, keep it simple stupid for end user compliance
- Why URI setup as is-->Recommended by ID.org, W3C HLSIG recommendations, established in RDF2Bio, easy content-negotiable dereferencing ...
- Why implementing semantic (labels) IDs rather than numeric IDs-->enable easy human compliant Sparql querying, increase readability.
- Why terminology/ontol selection & reuse-->Established ones, usage abundance, suitable semantics, domain applicability, ...

- Why technology/Implementation-->Virtuoso is free, widely used and expertise is hence abundant. It also comes with a comprehensive frontend and SPARQL endpoint.
- What Frontend--> Sparql query for batch querying and lists as expected results. A Linked Data Browser for on-the-fly content browsing along single tracks/link paths
 - What Visualizers? <http://www.w3.org/wiki/VisualizingRDF> , <http://semweb.salzburgresearch.at/apps/rdf-gravity/index.html>

Virtuoso and SPARQL endpoint setup

3 Example queries are installed on Virtuoso running on balzer.ipb-halle.de

Sources to Build are on balzer:/vol/debian/virtuoso-opensource

Web access via <http://balzer.ipb-halle.de:8890/>

The triplestore currently consists of a few thousand triples from 800 mass bank records (IPB and UFZ). The 800 record limit was set to alleviate performance problems with rdf.

Sparql-Queries can now be stated via "Linked Data" -> "SPARQL" (for saved query press "Edit")

Example Sparql queries & output

A) SPARQL QUERY

[/DAV/home/dba/query1.xml](#) Query for "Record link" and "Substance" where record has a certain peak (CQ 1)

#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 mbcodes:describes ?a.
```

```
  ?r mbcodes:hyperlink-record ?relink.
```

```
  ?a mbcodes:has-output ?s.
```

```
  ?s mbcodes:identifies ?c.
```

```
  ?c mbcodes:name ?substance.
```

```
  ?s mbcodes:contains-peak ?p.
```

```
  ?p mbcodes:has-mz "147.044"
```

```
}
```

RESULT

relink	substance
http://www.massbank.jp/SVN/OpenData/record/IPB_Halle/PB000122.txt	Naringenin

http://www.massbank.jp/SVN/OpenData/record/IPB_Halle/PB000167.txt	Kaempferol
http://www.massbank.jp/SVN/OpenData/record/IPB_Halle/PB005711.txt	Biochanin A

...

B) SPARQL QUERY

</DAV/home/dba/query2.xml> **Return all weblinks to records that have a substance whiches name contains "Naringenin"**

#This query outputs the links to all records which have a Substance #whose name contains "Naringenin"

```
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.
  FILTER regex(?substance, "Naringenin")
}
```

RESULT

relink	substance
http://www.massbank.jp/SVN/OpenData/record/IPB_Halle/PB000122.txt	Naringenin
http://www.massbank.jp/SVN/OpenData/record/IPB_Halle/PB000126.txt	Naringenin chalcone
http://www.massbank.jp/SVN/OpenData/record/IPB_Halle/PB000125.txt	Naringenin
http://www.massbank.jp/SVN/OpenData/record/IPB_Halle/PB000128.txt	Naringenin chalcone

C) SPARQL EXAMPLE

[/DAV/home/dba/query3.xml](#)

Return all weblinks to records that describe an MS2-Experiment

```
#Returns all records which contain an MS2-Experiment
#This returns the record entities, NOT the hyperlinks
#that can be done by replacing "?r" with "?reclink" before the "WHERE"
SELECT ?r
WHERE {
  ?r mbcodes:describes ?a.
  ?r mbcodes:hyperlink-record ?reclink.
  ?a mbcodes:ms-type "MS2"
}
```

RESULT

r
http://www.ipb-halle.de/ontologydata/record:PB000470
http://www.ipb-halle.de/ontologydata/record:PB000309
http://www.ipb-halle.de/ontologydata/record:PB000307

...

Abstracted (general) workflow for creating linked Data resources

We are currently working on a reusable domain independent flow scheme that guides LOD creators in setting up their RDF resources. This is partly inspired by <http://www.w3.org/2001/sw/hcls/notes/hcls-rdf-guide/>, but will be overhauled and presented in a graphical intuitive way.

- Review external existing RDF resources to re-use and integrate with
- Define Use case: Determines scope
- Define Competency Questions: Defines domain dependent content links
- Select the data sources or portions thereof to be converted to RDF in the case where re-use is no option..
 - *Identify the items of interest* in your domain, the things whose properties and relationships we want to describe
 - Agree on which items should be URI and which stay literals, e.g. float or string values
 - Identify persistent HTTP URIs for information & non-information resources (use hashed URIs here)
 - Choose your robust namespace
 - Use <http://identifiers.org/>

- Agree on Mime types you want to provide for HTTP content-negotiated additional representations presented upon dereferencing the URI, e.g. HTML in addition to RDF/XML
- Generate RDF model
 - First sketch handwritten graph models for all modules/namespaces envisioned, with links/edges between resources needed to answer CQs
 - Add core predicates/edges/relations
 - incl inverses/backlinks, and NS where to take them from
 - Add a 3-node-spanning link along multiple namespaces, i.e. to show nested queries & how URIs are used as primary key to pass information along
 - List of what essential literals in resource modules can become robust URIs
 - Must be translatable with high confidence (e.g. SpeciesLabel="brassicaceae" → NCITax:ID723345)
 - URI examples in accordance to ID.org (dereferenceable via content negotiation)
 - for all key nodes/predicates in all NS modules
 - existing ones AND own ones (get server name here for HTTP url to serve own NS)
 - To align ChemicalIDs use <https://www.ebi.ac.uk/unichem/>
- Generate ontology defining the formal semantics of the RDF model.
- Generate example RDF triples
 - in turtle or RDF/XML syntax and along the CQ scope, i.e.
 - 1. using own NS-URI (Massbank) to literal (e.g. Mass value as float?)
 - 2. using own NS-URI (Massbank) to own NS-URI (Massbank)
 - 3. using own NS-URI (Massbank) to established external resource URI (e.g. ChEBI from BioToRDF or Bioportal)
- Publish the RDF data as Linked Data or through SPARQL endpoint
 - E.g. set up Virtuoso server with endpoint & configure it.
- Agree on Information from different sources that merges in naturally & allows synergistic insights (context enrichment)
- Set RDF links between data from different external sources
- Create Semantic Web applications using the published data.
- Add example SPARQL queries along use cases in a) human readable form and b) in annotated SPARQL, e.g. turtle syntax
 - Queries should leverage on the example triples
 - store and document accompanying result data sets.
 - Build and add query library
- Write Documentation
- Register and make your LOD known with sem web crawlers/tools/Websites

Best Practice Documents for Linked Data that we adhere to

<http://wifo5-03.informatik.uni-mannheim.de/bizer/pub/LinkedDataTutorial/>

<http://www.w3.org/2001/sw/hcls/notes/hcls-rdf-guide/>
<http://latc-project.eu/best-practices>
https://docs.google.com/document/d/1U44mbDeJGE-_bfL16lcOx3eaeBkTnDOHDSjkYF1XHMc/edit#heading=h.cw4nqjg7epgt
<http://www.w3.org/DesignIssues/LinkedData.html>
<https://www.cambridgesemantics.com/semantic-university/learn-semantic-web-design-patterns>

Best practices for Life Science Linked Data:

<http://www.sciencedirect.com/science/article/pii/S1570826812000376>

On IDs as URIs:

<http://www.obofoundry.org/id-policy.shtml>

Literature references

- <http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf>
- Erwin Folmer, Paul Oude Luttighuis and Jos van Hillegersberg: *Do semantic standards lack quality? A survey among 34 semantic standards*. *Electronic Markets*. Volume 21, Number 2 (2011), 99-111. DOI: [10.1007/s12525-011-0058-y](https://doi.org/10.1007/s12525-011-0058-y).
- Alejandra Gonzalez-Beltran, Eamonn Maguire, Susanna-Assunta Sansone, Philippe Rocca- Serra. *linkedISA: semantic representation of ISA-Tab experimental metadata*. *BMC Bioinformatics* 2014, *in press*.

SemWeb & LOD Tutorials

<http://www.cambridgesemantics.com/semantic-university/getting-started>
<https://github.com/JoshData/rdfabout/blob/gh-pages/intro-to-rdf.md#>
<https://www.cambridgesemantics.com/semantic-university/rdfs-vs.-owl/>
<http://www.cambridgesemantics.com/semantic-university/learn-sparql>

Useful hyperlinks

abundantly used endpoints: <http://www.listnerd.com/list/best-life-science-sparql-endpoint-off-all-time>

Virtuoso: <http://virtuoso.openlinksw.com/white-papers/>

Tools

General

<http://semanticweb.org/wiki/Tools>
<http://www.w3.org/wiki/SemanticWebTools>
<http://wifo5-03.informatik.uni-mannheim.de/latc/toollibrary/>
<http://haystack.csail.mit.edu/>

RDFizers

<http://simile.mit.edu/wiki/RDFizers>

From XML & XHTML to RDF

[GRDDL](#) - Gleaning Resource Descriptions from Dialects of Languages. A mechanism for getting RDF data out of XML and XHTML documents using explicitly associated transformation algorithms, typically represented in XSLT.

Cross-resource Query Federation in SPARQL

<http://arxiv.org/pdf/1306.1723.pdf>

<https://informatik.hu-berlin.de/forschung/gebiete/wbi/research/publications/2008/DARQ-FINAL.pdf>

Semantic Web Search Engines

<http://www.w3.org/wiki/TaskForces/CommunityProjects/LinkingOpenData/SemanticWebSearchEngines>

Open Linked Data Browsers

<http://www.w3.org/wiki/TaskForces/CommunityProjects/LinkingOpenData/SemWebClients>

<http://notes.3kbo.com/linked-data-browsers>

<http://sites.tufts.edu/liam/2014/03/22/browsers/>

<http://ode.openlinksw.com/>

<http://munkapad.sztaki.hu/lodmilla/>

http://simile.mit.edu/wiki/Piggy_Bank

Validators

RDF Validator

<http://www.w3.org/RDF/Validator/>

LOD Validators

VAPOUR: <http://validator.linkeddata.org/vapour>

Checks whether semantic web data is correctly published according to the current best practices, as defined by the [Linked Data](#) principles, the [Best Practice Recipes](#) and the [Cool URIs](#).

(Outdated?: <http://tripletest.sourceforge.net/>)