NFDI4Chem

ENHANCE
YOUR
DATA.

TIB

dct:title

Breakout Session II: Hands on Data Annotation using Ontologies

Creating a prototype knowledge graph from NMR spectroscopy research data

dct:identifier DOI:10.5281/zenodo.7050763

dct:creator ORCID:0000-0002-1595-3213

ro:'participates in' 1st Ontologies4Chem Workshop 2022 – Ontologies for chemistry (day 2)

- dct:identifier DOI:10.25798/frnp-sn04
- sdo:duration 2022-09-08T16:50:00+02:00/PT10M

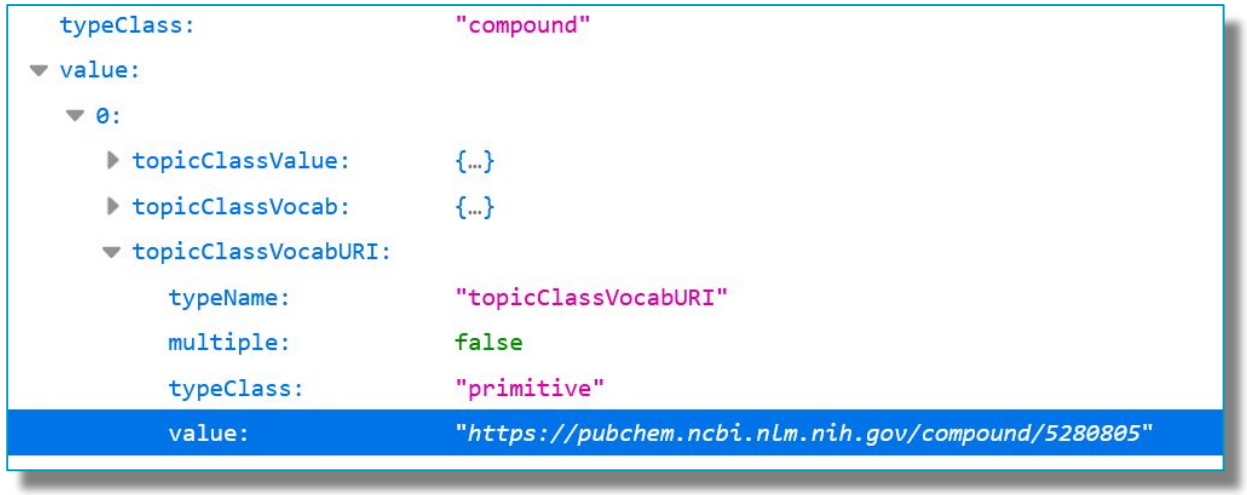
Creating knowledge graphs from research data
the preconditions - research data formats

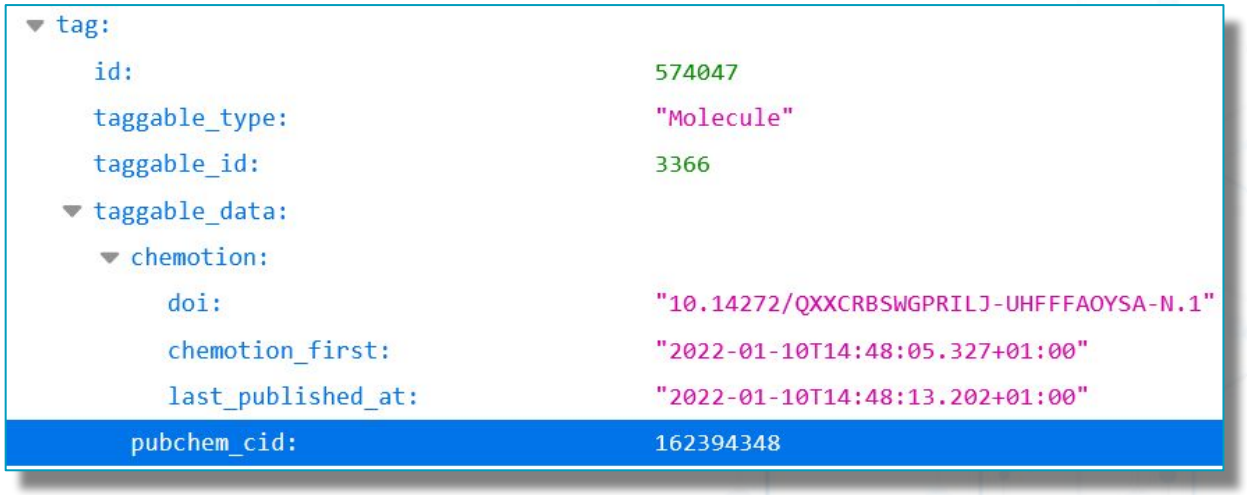
- research data is stored in multiple file formats
  - domain specific formats
    - vendor specific or open file formats

Format	Data type	Maintainer	Parent Format	Specification
mzML	Mass spectrometry	HUPO/ PSI	XML	open
NMReDATA	NMR	NMReDATA Initiative	SDF	open
JCAMP-DX	multiple	IUPAC	ASCII, Text	open

Creating knowledge graphs from research data
the preconditions - metadata schemata

- there are different metadata schemata
  - e.g. DataCite's schema, schema.org, Dublin Core Terms and repository specific schemata
    - often not domain specific / fine grained enough
    - if domain specific usually repository specific





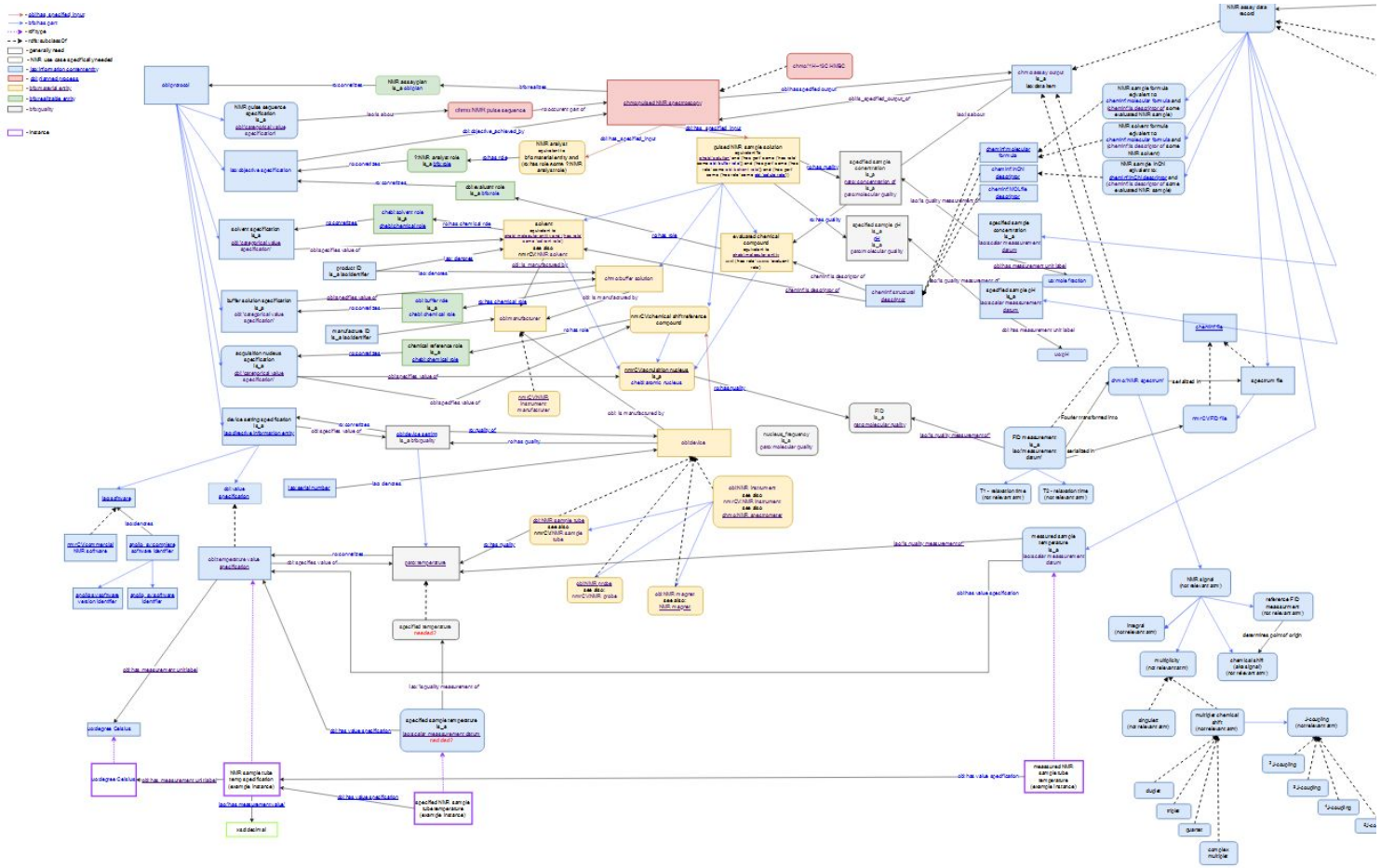
Creating knowledge graphs from research data
the preconditions - domain specific metadata

- it is hard to query & aggregate data from different sources
  - need to know provider specific metadata schemata
  - or, have a true domain specific standard schema
    - like IUPAC FAIRspec finding aid ?

Creating knowledge graphs from research data
the preconditions - suitable ontologies already exist

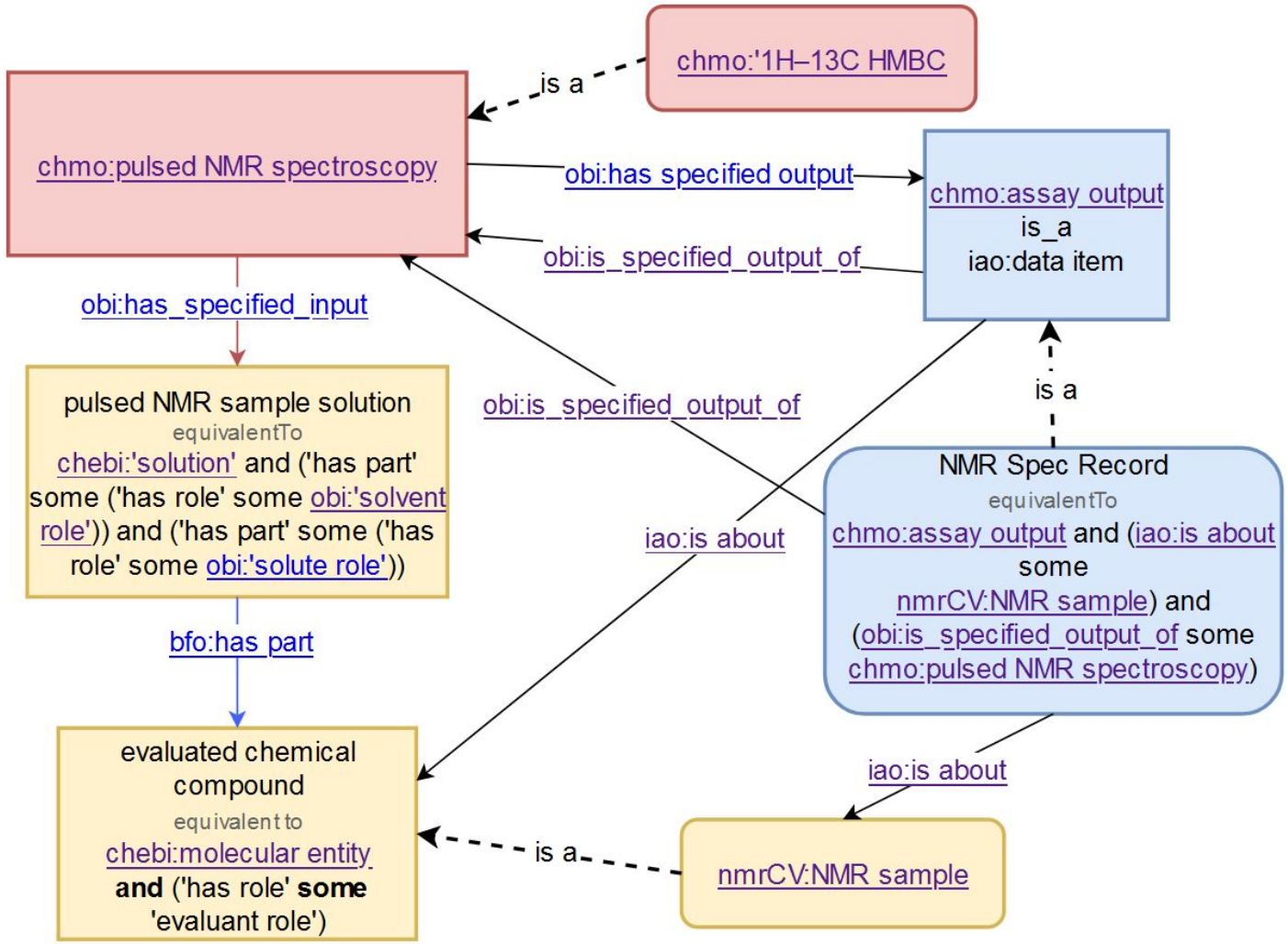
- knowledge graphs could be an alternative
  - make federated queries with SPARQL
  - only need to know about the used ontology terms
    - many of which are already commonly used
  - fosters interdisciplinary work
- proof of concept
  - make knowledge graphs from existing datasets
  - make federated queries on them with SPARQL

Creating knowledge graphs from research data
Step 1: define a terminology box (TBox)



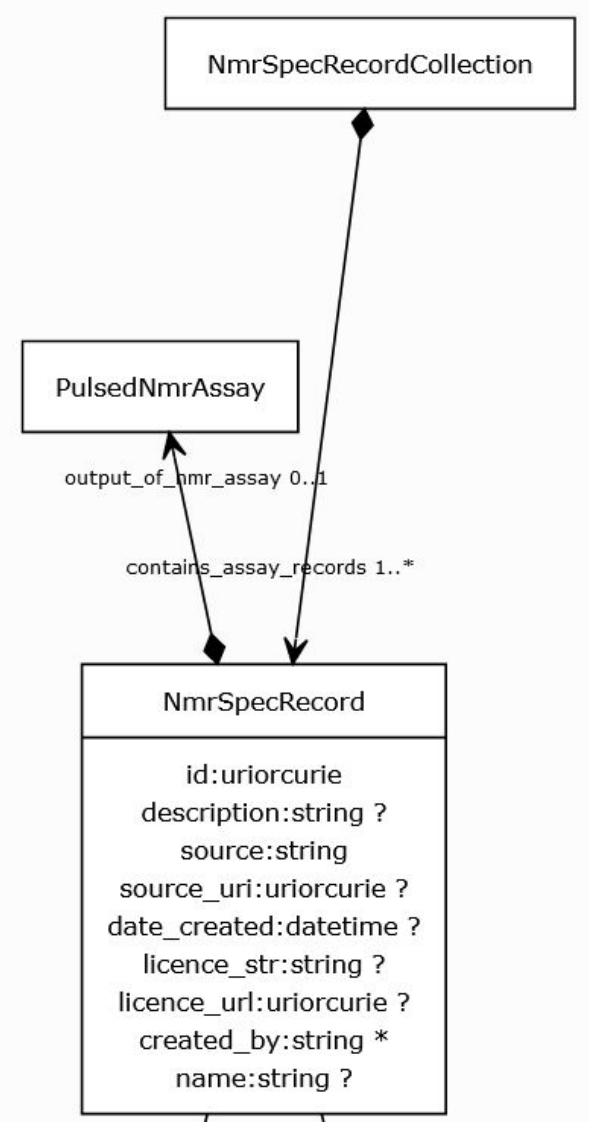
- finding the right terms to describe NMR spectroscopy
- reusing as many terms as possible for interoperability
- we found most in:
  - CHMO
  - nmrCV
  - CHEBI
  - CHEMINF
  - OBI
  - IAO

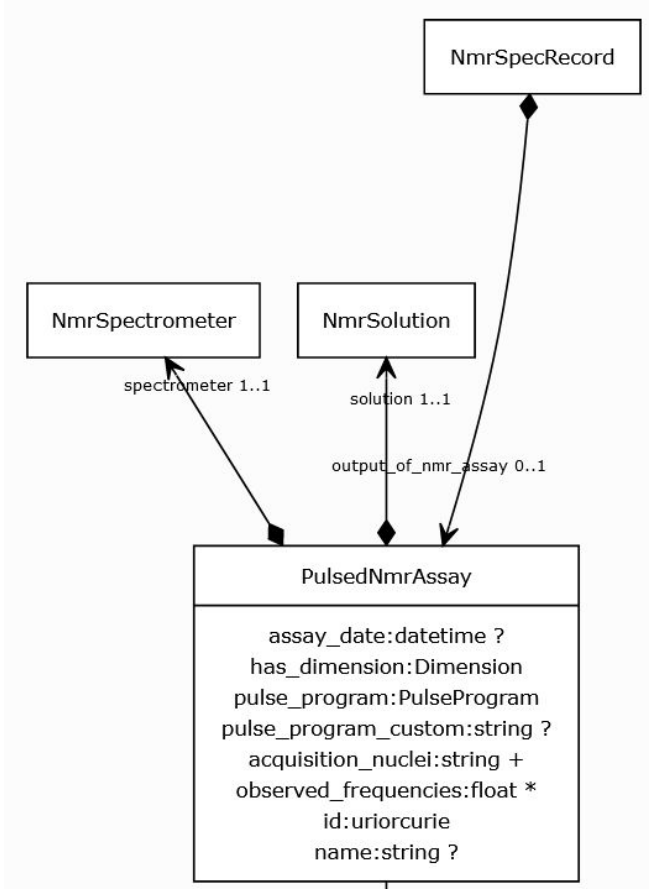
Creating knowledge graphs from research data
Step 1: define a terminology box (TBox)



- need to define new entities
  - in existing ontologies
  - one of our later TODOs
  - involves discussion with domain experts
  - semantically more precise

Creating knowledge graphs from research data
Step 2: define the shape of the knowledge graph

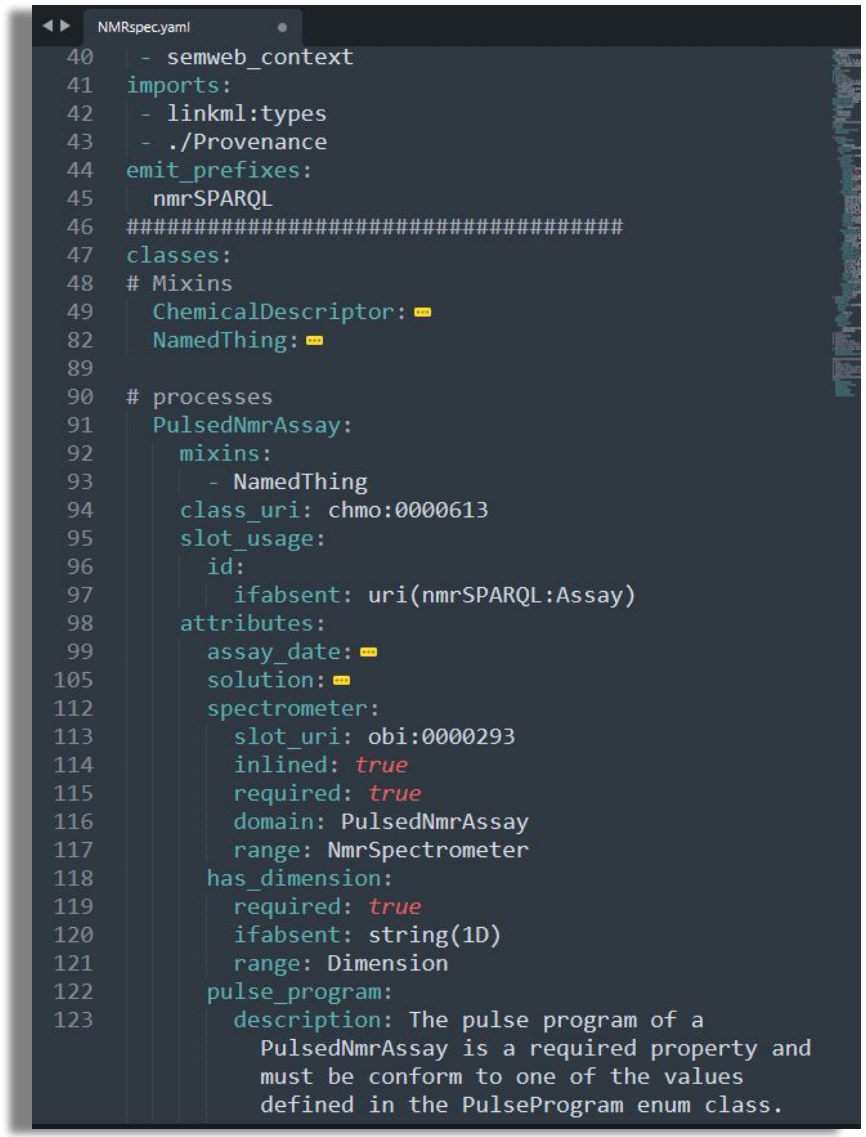


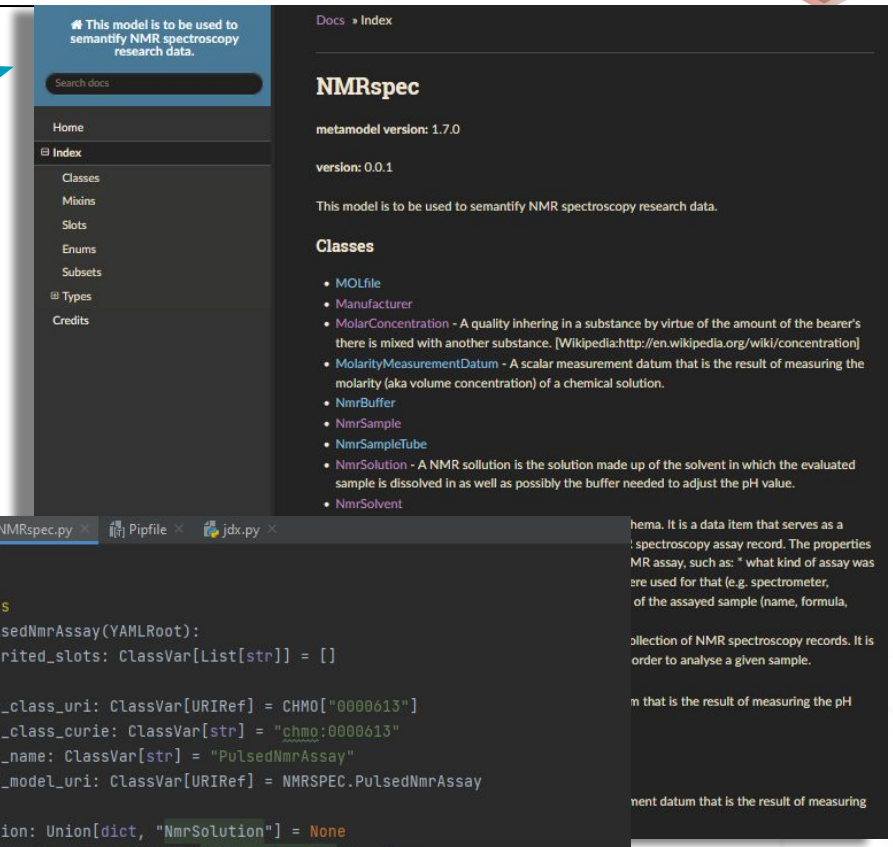


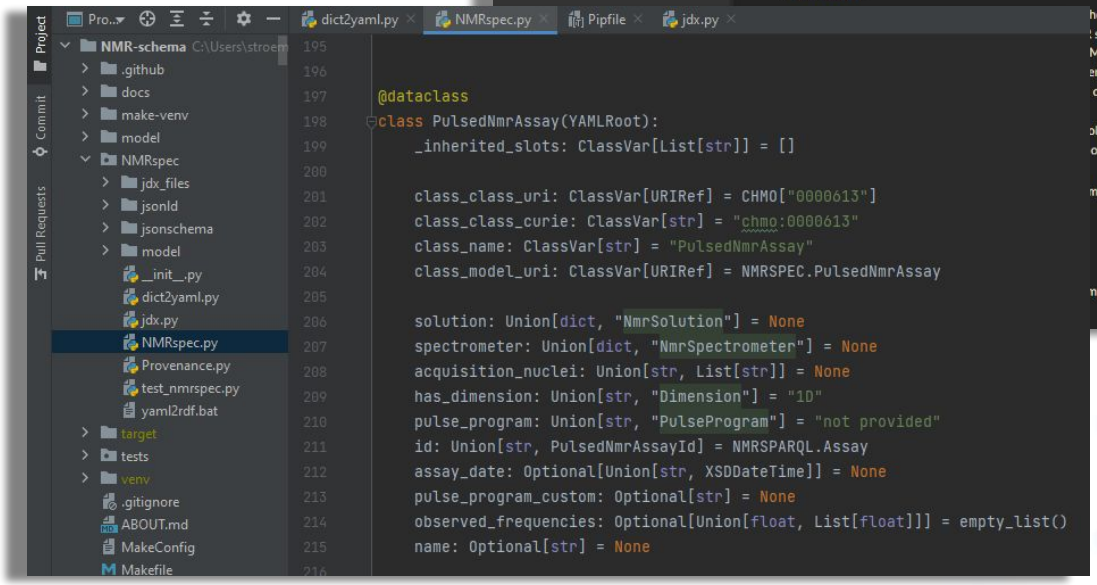
- https://stroemphi.github.io/NMRspec/
- using https://linkml.io framework
- prototypical → not complete
- focus on pulsed NMR spectroscopy
  - only links to broader assay context

Creating knowledge graphs from research data

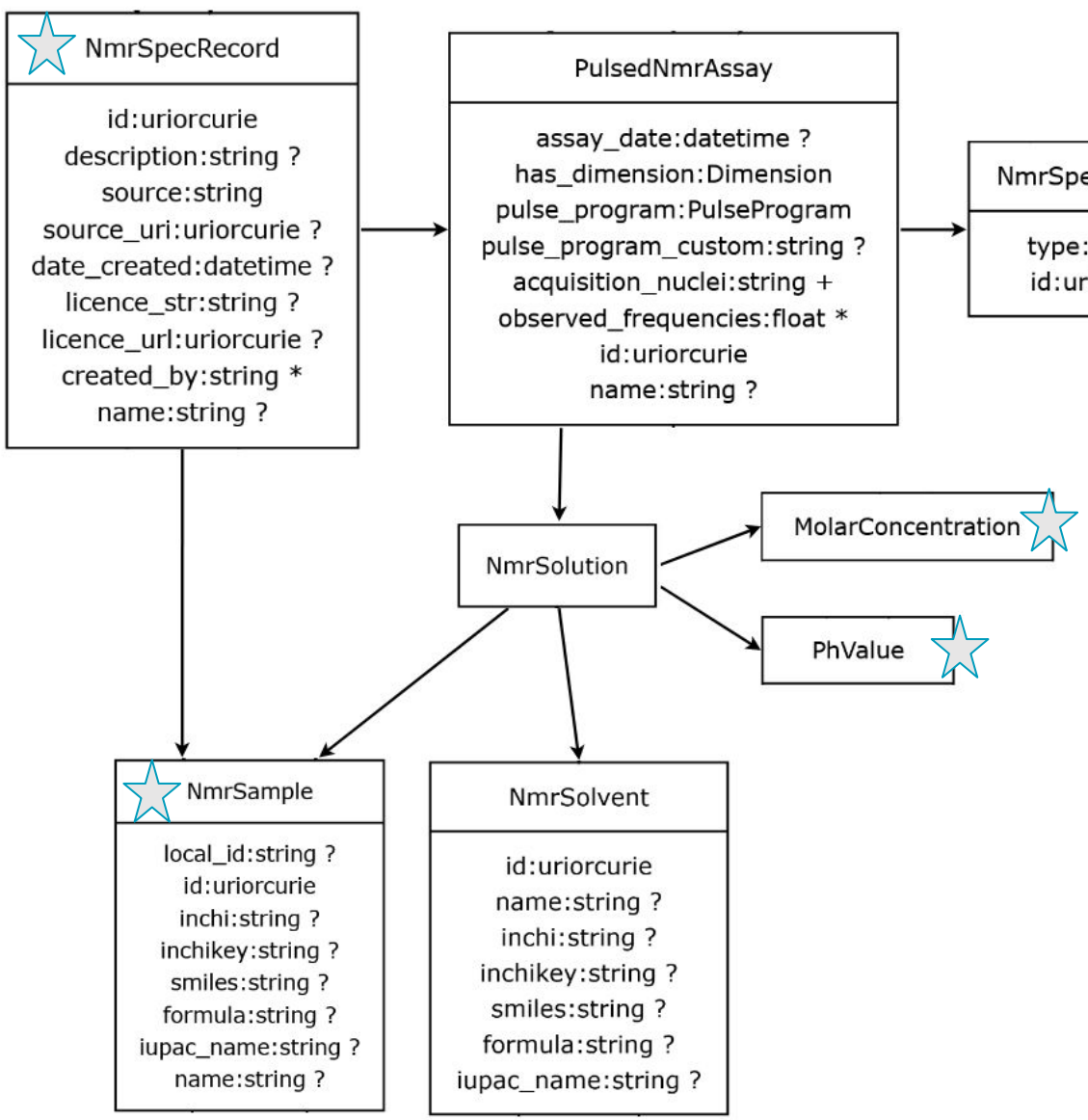
Step 2: define the shape of the knowledge graph





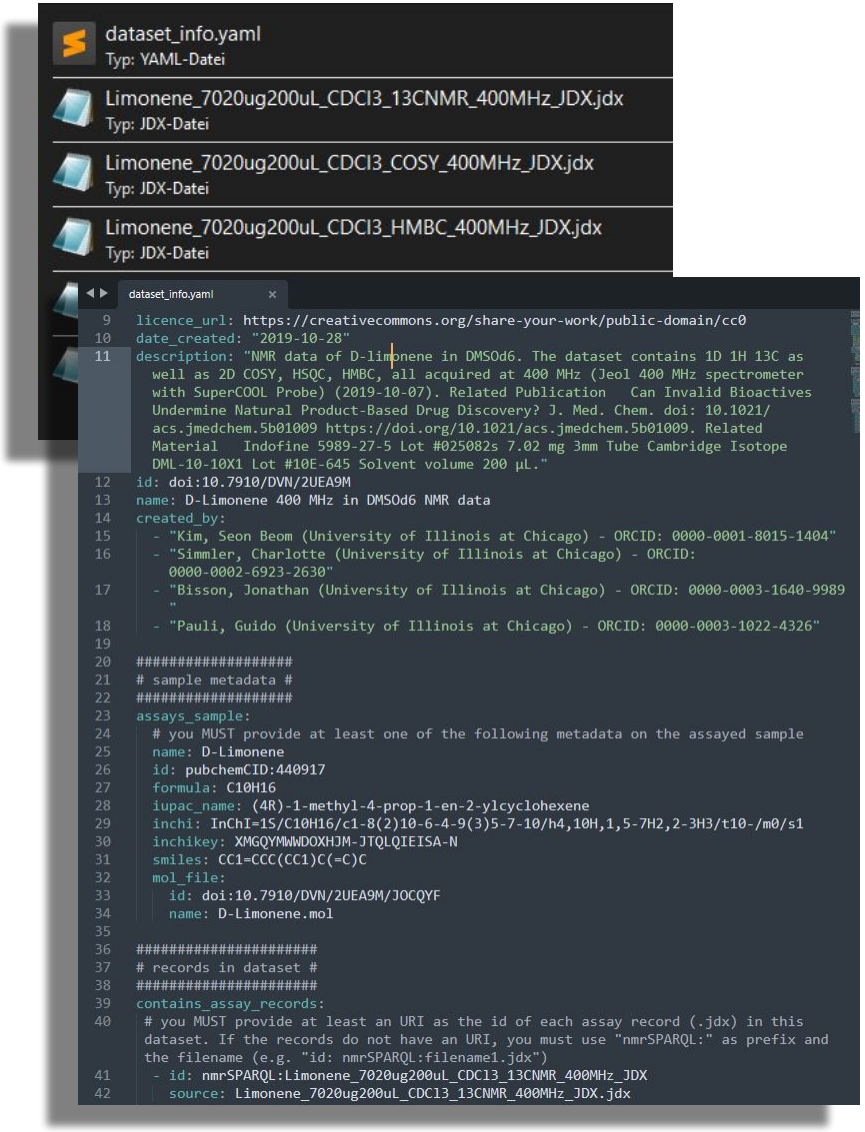


Creating knowledge graphs from research data
Step 3: parsing values from JCAMP-DX



- JCAMP-DX datasets as source for NMR
  - around for long and open
  - Python parser already exists
  - unfortunately many dialects :(
  - contains only limited metadata
    - ☆must come from other sources
    - assumed to be auto-generated in ELNs
- still hard to get many .jdx files
  - we used Havard Dataverse, Chemotion, OSDB

Creating knowledge graphs from research data
Step 3: parsing values from JCAMP-DX



- dataset_info.yaml with metadata not in JDX files
  - dataset provenance (source, creators, PID, …)
  - metadata of the assayed sample (InChI, InChIKey, SMILES, …)
- dict2yaml.py
  - parse JDX files in batch
  - create one NMR record collection as YAML
    - each JDX is one NMRSpecRecord
- yaml2rdf.bat
  - covert the YAML to TTL

Creating knowledge graphs from research data
Step 4: load the knowledge graphs into a triple store

- NMR research metadata expressed in triple statements

```
<https://doi.org/10.7910/DVN/F34GVS/KZOQZU/1HNMR_jdx> a ns1:NmrSpecRecord ;
    ns2:OBI_0000312 <https://doi.org/10.7910/DVN/F34GVS/KZOQZU/1HNMR_jdx/Assay> ;
    dc:source "1HNMR_jdx.jdx" .

<https://doi.org/10.7910/DVN/F34GVS/KZOQZU/1HNMR_jdx/Assay> a ns2:CHMO_0000613 ;
    ns2:OBI_0000293 <https://doi.org/10.7910/DVN/F34GVS/KZOQZU/1HNMR_jdx/Solution>,
        <https://doi.org/10.7910/DVN/F34GVS/KZOQZU/1HNMR_jdx/Spectrometer> ;
    ns1:acquisition_nuclei "1H" ;
    ns1:has_dimension "1D" ;
    ns1:observed_frequencies "399.78219837825"^^xsd:float ;
    ns1:pulse_program "NMR" .
```

SPARQL Queries

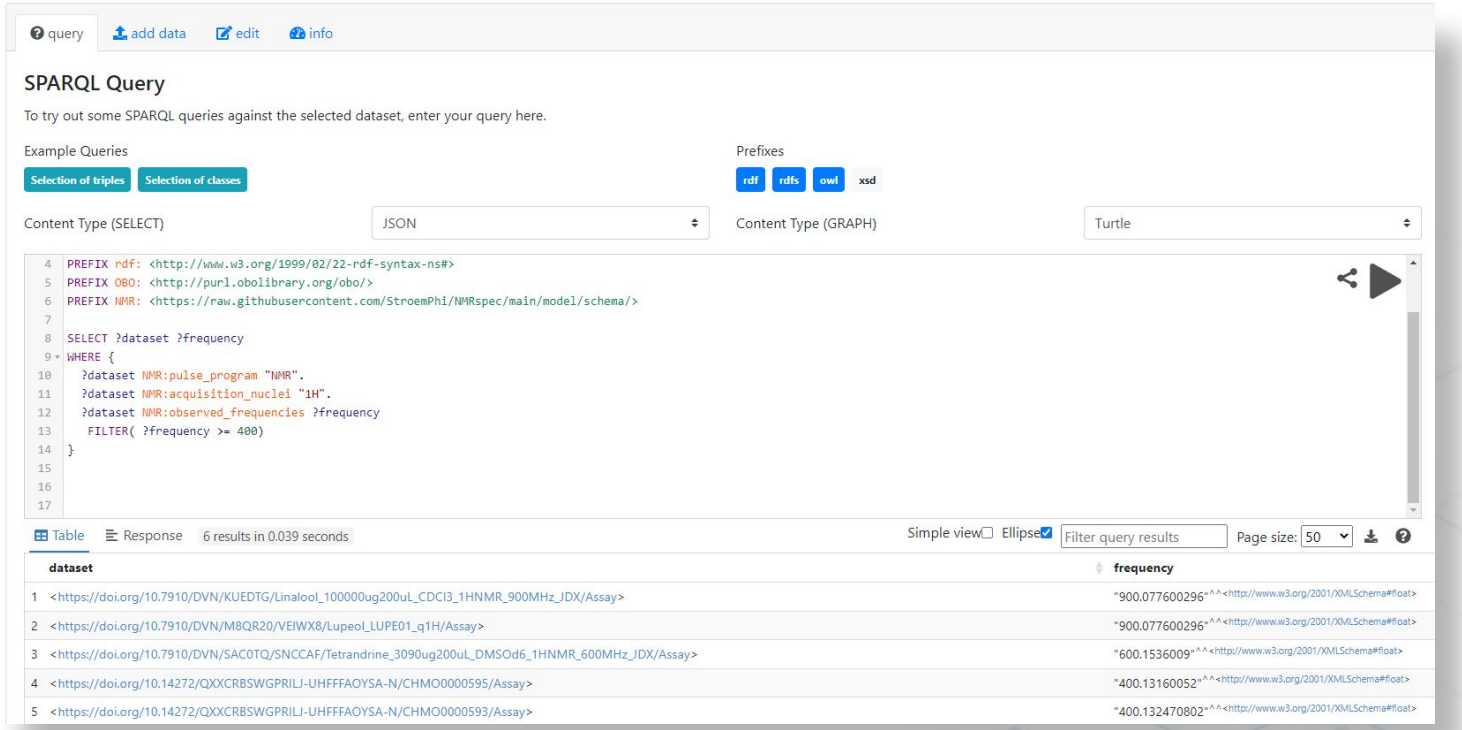
Show me only 1H NMR Spectra with a resolution higher 400 MHz

PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

PREFIX OBO: <http://purl.obolibrary.org/obo/>

PREFIX NMR: <https://raw.githubusercontent.com/StroemPhi/NMRspec/main/model/schema/>

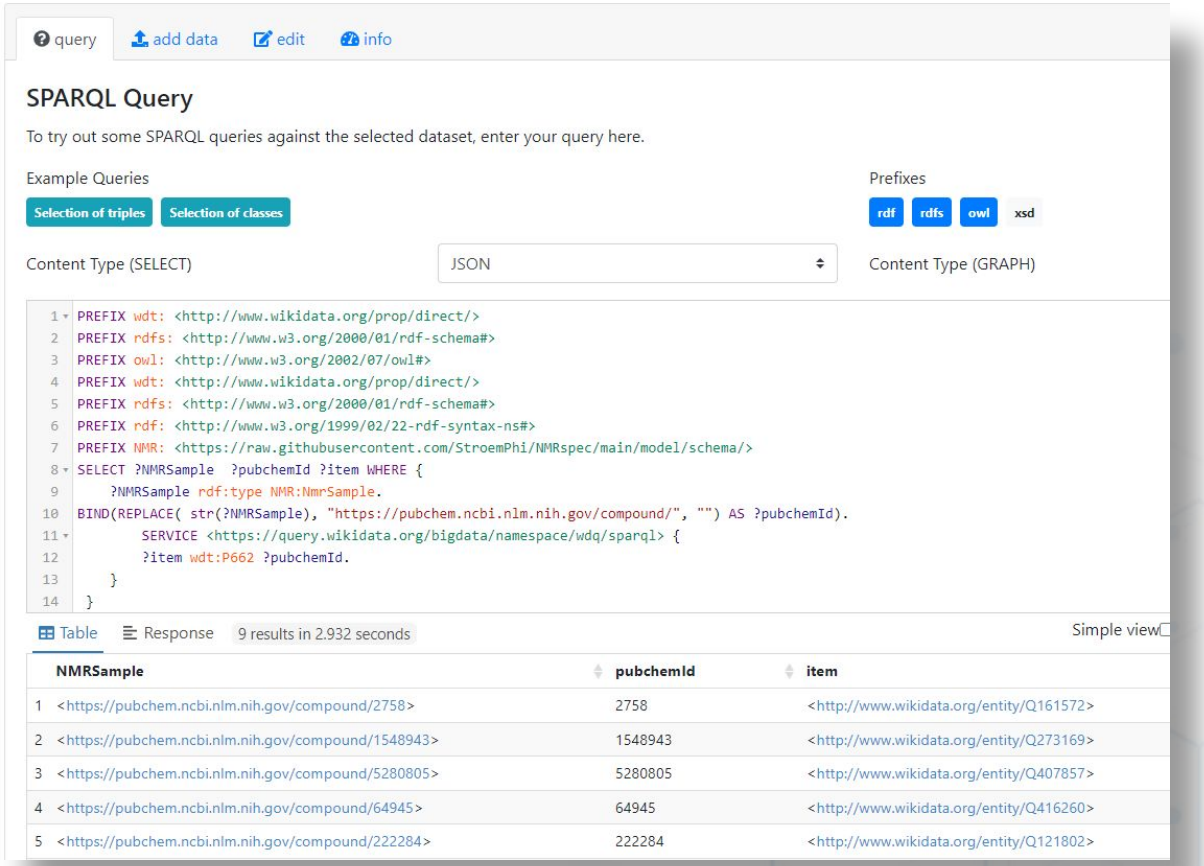
```
SELECT ?dataset ?frequency
    WHERE {
     ?dataset NMR:pulse_program "NMR".
     ?dataset NMR:acquisition_nuclei "1H".
     ?dataset NMR:observed_frequencies ?frequency
      FILTER( ?frequency >= 400)
    }
```



Federated SPARQL Queries

Federated query across Research Data Triplestore and Wikidata based on PubChemId

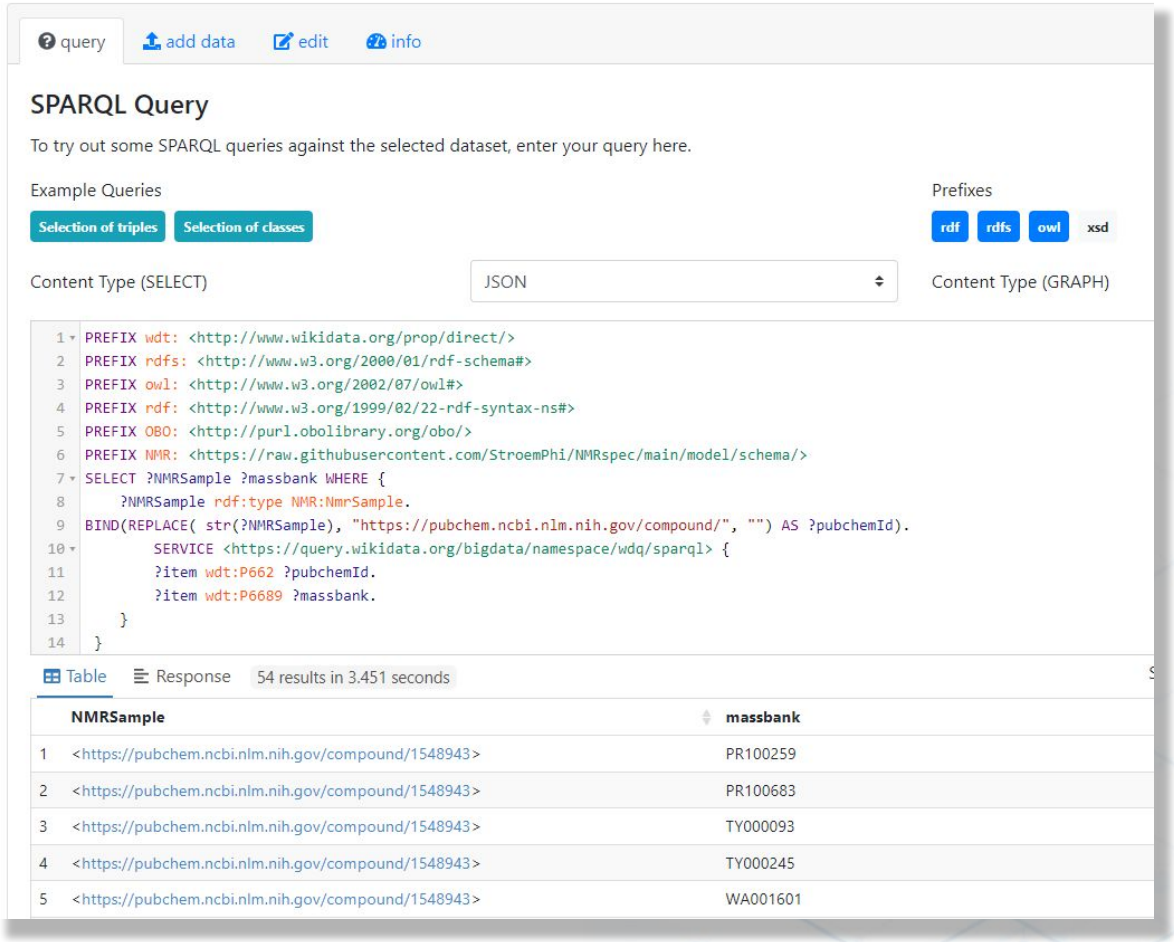
```
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX NMR: <https://raw.githubusercontent.com/StroemPhi/NMRspec/main/m
SELECT ?NMRSample  ?pubchemId ?item WHERE {
        ?NMRSample rdf:type NMR:NmrSample.
        BIND(REPLACE( str(?NMRSample),
        "https://pubchem.ncbi.nlm.nih.gov/compound/", "")
        AS ?pubchemId).
        SERVICE <https://query.wikidata.org/bigdata/namespace/wdq/sparql>
        {
                ?item wdt:P662 ?pubchemId.
        }
}
```



Federated SPARQL Queries

Find matching entries in further data repository: Like finding a corresponding Mass Spectra in MassBank

```
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX NMR: <https://raw.githubusercontent.com/StroemPhi/NMRspec/main/
SELECT ?NMRSample  ?pubchemId ?item WHERE {
        ?NMRSample rdf:type NMR:NmrSample.
        BIND(REPLACE( str(?NMRSample),
                "https://pubchem.ncbi.nlm.nih.gov/compound/", "")
                AS ?pubchemId).
        SERVICE <https://query.wikidata.org/bigdata/namespace/wdq/sparql>
        {
                ?item wdt:P662 ?pubchemId.
                ?item wdt:P6689 ?massbank.
        }
}
```



NFDI4Chem

ENHANCE
YOUR
DATA.

TIB

Thank you