Leveraging knowledge graphs for efficient use of laboratory instrument data

Spring 2024 Allotrope Connect


Jindra Mynarz


Max Skoryk

MSD Czech Republic

2024-05-15

Knowledge graphs

Knowledge graph is a dataset representing a **part of the real world in a graph database** (sometimes using formal semantics).

It can store **both data and ontologies** and allows querying across them. The data is thus **self-describing**.

Ontologies provide an additional leverage for **semantic queries**, such as for query expansion via ontological relations.

Using knowledge graphs **reduces data friction** involved in combining data from multiple heterogeneous data sources.

We argue that knowledge graphs can provide **analysis-ready data** and make the use of laboratory instrument data **efficient**.

Allotrope Simple Models: ASM patterns

ASM patterns define the standard ways how to interpret ASM data using the semantics of **Allotrope Foundation Ontologies** (AFO).

In turn, they can guide an **ASM JSON schema-aware transformation** of data from ASM files to knowledge graphs based on the W3C's **semantic web standards**.

Quantity datum

ASM JSON Schema:

```
{
  "$asm.pattern": "quantity datum"
}
```

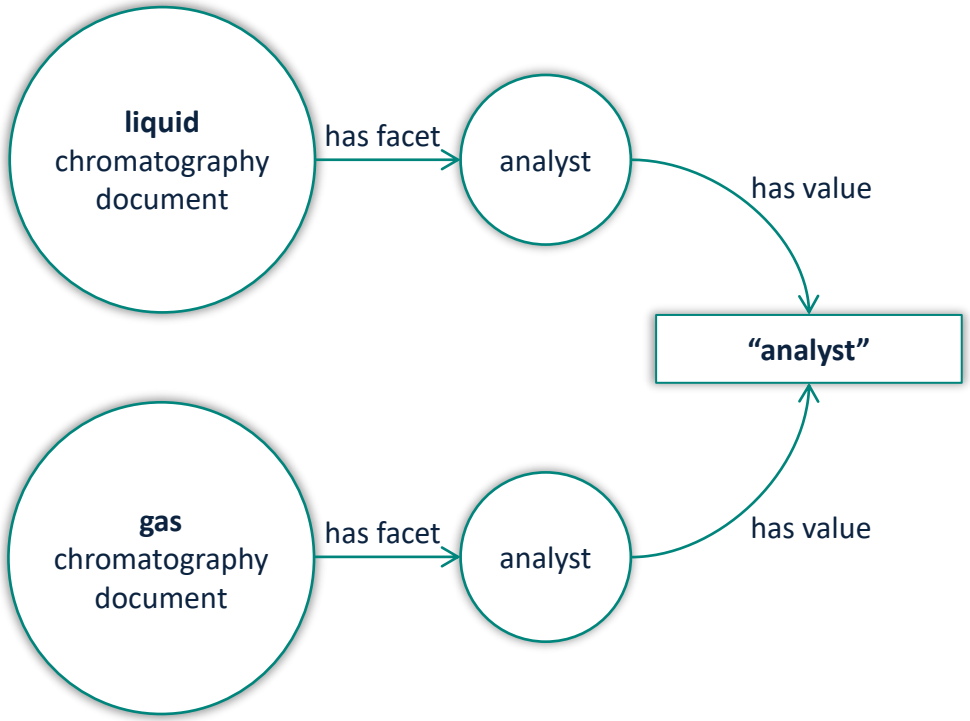
ASM JSON:

```
{
  "$CLASS_LABEL": {
    "value": "$VALUE",
    "unit": "$UNIT_SYMBOL"
  }
}
```

RDF/Turtle:

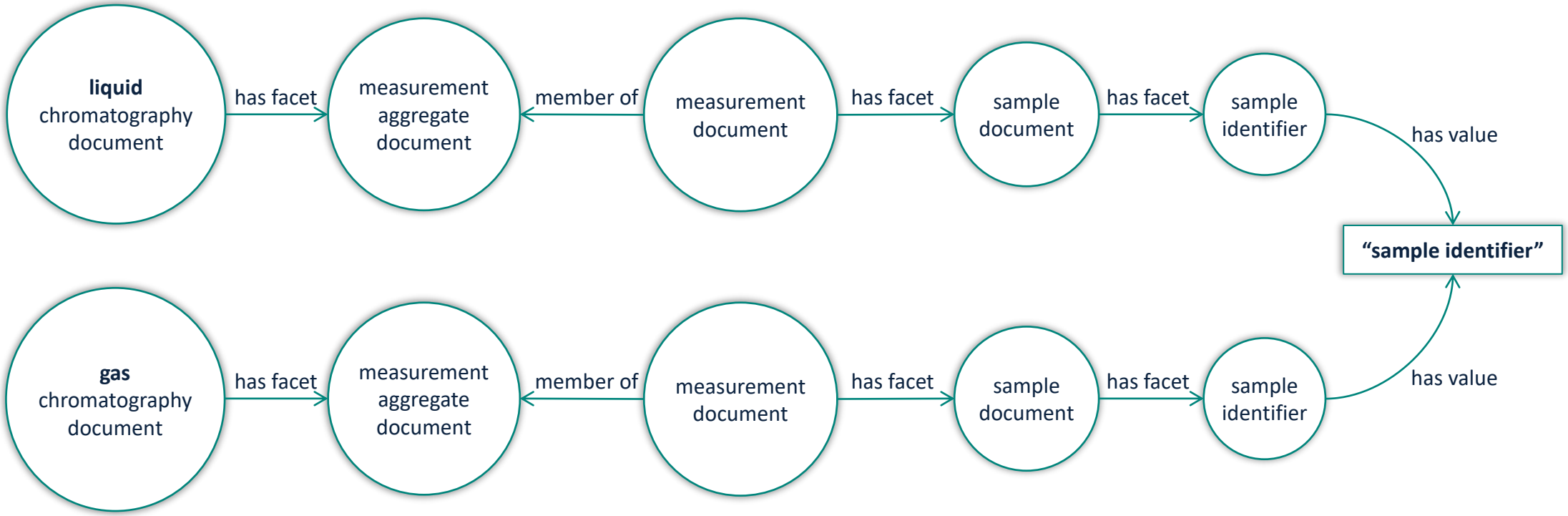
```
[] a <$CLASS> ;
  qudt:numericValue "$VALUE"^^<$DATA_TYPE> ;
  qudt:unit <$UNIT_CLASS> .

<$CLASS> skos:prefLabel "$CLASS_LABEL"^^xsd:string .
<$UNIT_CLASS> qudt:symbol "$UNIT_SYMBOL"^^xsd:string .
```

Queries across laboratory instrument data #1



Knowledge graphs are stored in **graph databases** which allow **efficient joins** across multiple sources of data, such as laboratory instrument runs.

Queries across laboratory instrument data #2



Constructing knowledge graphs out of ASM files

In 2023 we ran a **proof-of-concept** creating a knowledge graph out of ASM files using a **schema-aware transformation of ASM files to data RDF** implementing the standard ASM patterns.

Each ASM file is stored in a separate **named graph** to preserve the data separation via ASM files as containers. This is important due to the **lack of global identifiers of instances** (i.e. IRIs) in ASM files (all instances are identified by blank nodes).

Use the right tool for the job: **ASM data cubes** are not loaded into the graph. However, the knowledge graph can serve as a data catalog navigating users to data cubes via metadata (i.e. FAIR's **findability**).

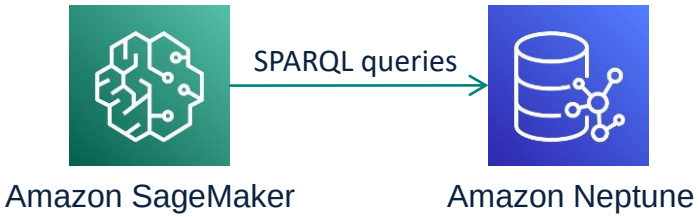
To aid readability of queries, we **alias opaque identifiers of AFO terms with human-readable names** via SPARQL prefixes.

Demo of a proof-of-concept

We constructed a knowledge graph out of a small subset of our laboratory instrument data from the chromatography domain. This included:

- Liquid chromatography
- Gas chromatography
- Fast protein liquid chromatography

The demo shows graph queries using a SageMaker notebook connected to an Amazon Neptune graph database.



Demo of a proof-of-concept: overview

Queries across multiple laboratory instrument runs: by laboratory instruments, manufacturers, sample identifiers, analysts

Self-describing data: labels and definitions from Allotrope Foundation Ontologies, hierarchy of peak facets, explicit units of measurement, semantic metadata of data cubes

Tracing the origin of data: source ASM files and raw instrument files

Validation of department-specific data quality rules: expected suitability pharmacopoeia settings

Wish list

Zero-copy data integration

Avoid the overhead of copying data**.** Fix the cause (data capture), not the symptoms. Laboratory instrument vendors should consider **native support of ASM**.

Improve quality control of Allotrope standards

Some Allotrope artifacts (e.g., ASM JSON schemas) contain errors, such as syntax errors, broken links, missing annotations (e.g., “$asm.pattern”). Each Allotrope’s user must resolve these errors to make the artifacts machine-readable. Improving quality control of Allotrope standards can help **share the cost of resolving these errors**.

Summary

We constructed a **knowledge graph of laboratory instrument data** out of ASM files. This PoC was successful and will be the basis for serving data to our scientific and data science communities.

Knowledge graphs allow **querying across multiple data sources and ontologies** in a single data space. This allows:

- **Semantic queries:** e.g., query expansion via ontological relations, such as class subsumption
- **Self-describing data**: e.g., queries can get definitions of ontology terms
- **Findability of non-semantic data**: e.g., navigating to data cubes via ASM metadata
- **Expressive queries**: e.g., testing compliance with internal rules on top of the ASM JSON schema rules
- **Machine-readable traceability**: e.g., for audit purposes
- **Query federation**: e.g., enrichment with contextual data