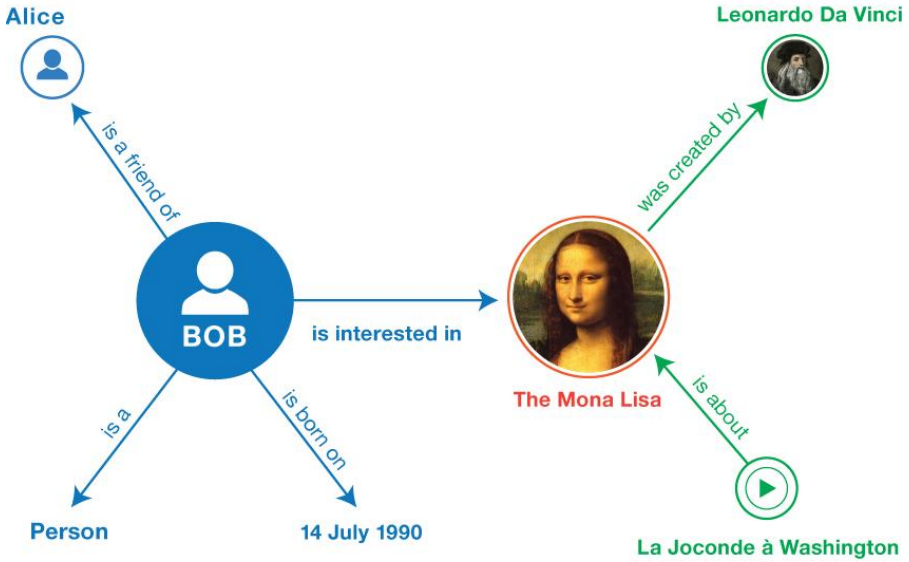
Leveraging HDF5 infrastructure by ADF to build an interoperable package & contextualized data in Pharma using semantic technology

Amnon Ptashek
Allotrope Foundation

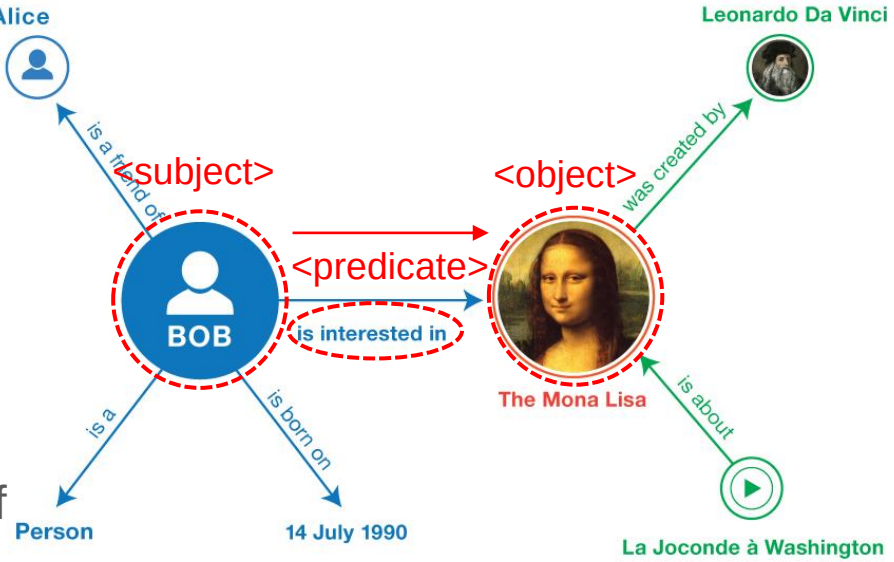
RDF Primer

- The ***RDF*** - Resource Description Framework is a framework for expressing knowledge about resources.
- Resources can be anything: devices, people, physical objects, and abstract concepts
- We can visualize knowledge as a connected graph, consisting of nodes and arcs



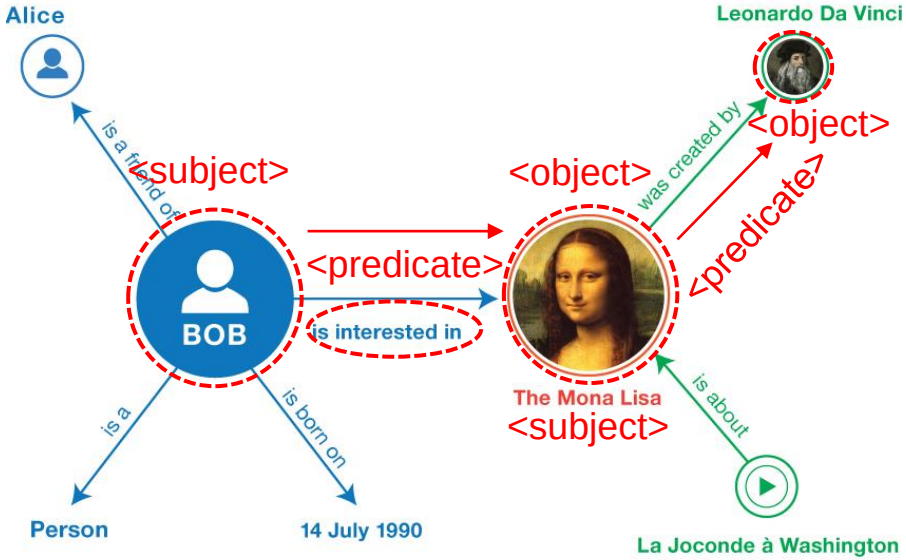
RDF Primer

- ***RDF*** statement expresses a relationship between two resources
- A statement always has the following structure called **triples**:
  <subject> <predicate> <object>
- The ***subject*** and the ***object*** represent the two resources being related; the ***predicate*** represents the nature of their relationship.
- The relationship is phrased in a directional way from ***subject*** to ***object*** and is called in ***RDF*** a ***property***



RDF Primer

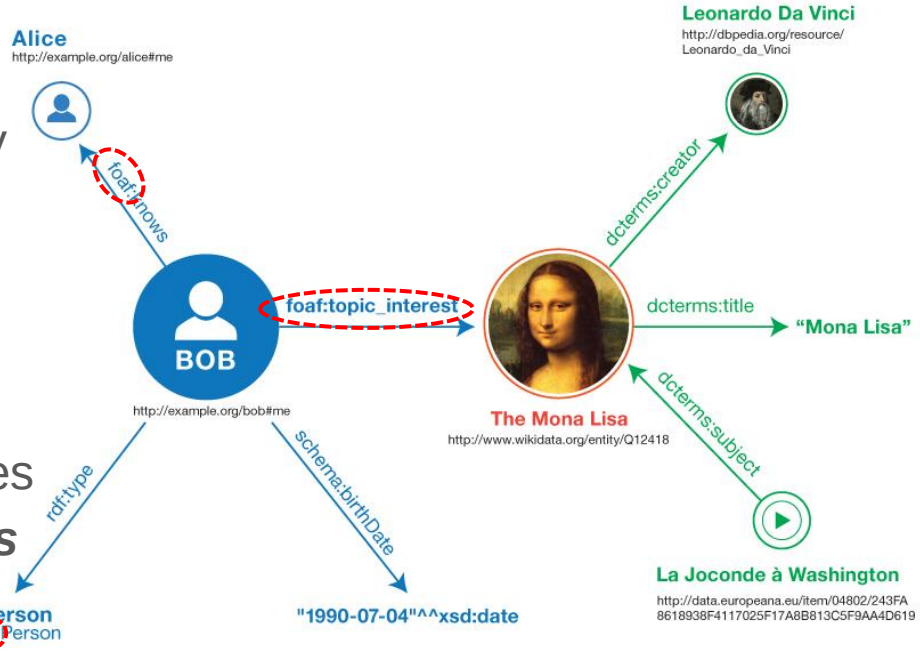
- This ability to have the same resource be in the ***object*** position of one triple and the ***subject*** position of another triple makes it possible to find connections between triples, which makes ***RDF*** very powerful



<Bob> <is a> <person>.
<Bob> <is a friend of> <Alice>.
<Bob> <is born on> <14th of July 1990>.
<Bob> <is interested in> <the Mona Lisa>.
<the Mona Lisa> <was created by> <Leonardo da Vinci>.
<the video 'La Joconde à Washington'> <is about> <the Mona Lisa>

RDF Primer

- ***IRI*** - International Resource Identifier are used to identify resources uniquely
- ***Objects*** can be ***literal***s that are basic values and not ***IRI***s
- ***IRI***s are typically used in combination with ***RDF vocabularies*** that provide semantic context about these resources
- There are universal ***RDF vocabularies*** such as "Friend of a Friend" (FOAF) vocabulary for describing social networks



http://xmlns.com/foaf/0.1/Person
http://xmlns.com/foaf/0.1/knows
http://xmlns.com/foaf/0.1/topic_interest

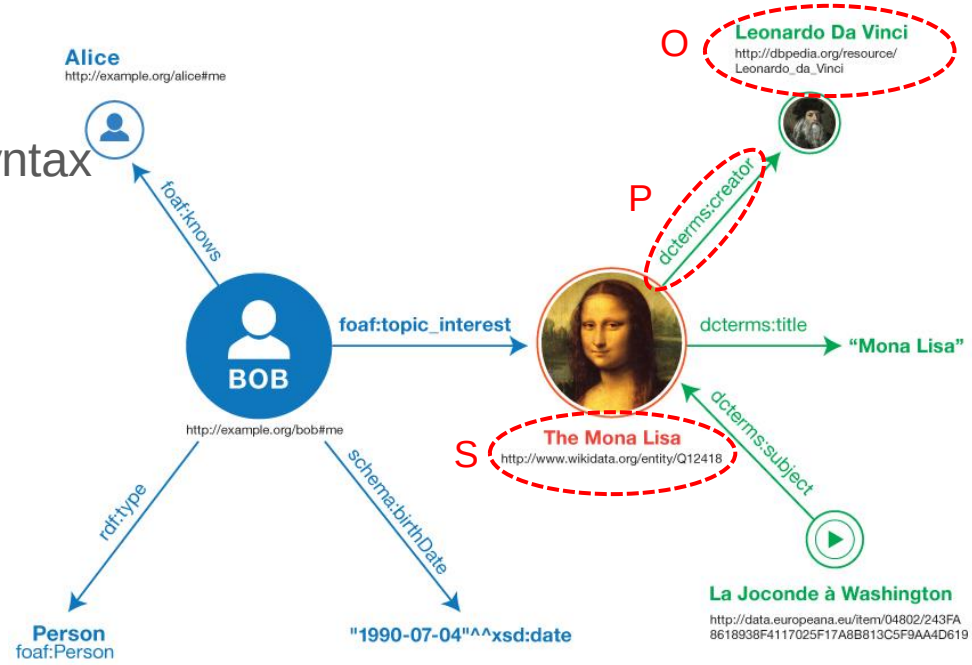
Namespace to Prefix mapping

foaf:Person
foaf:knows
foaf:topic_interest

RDF Primer

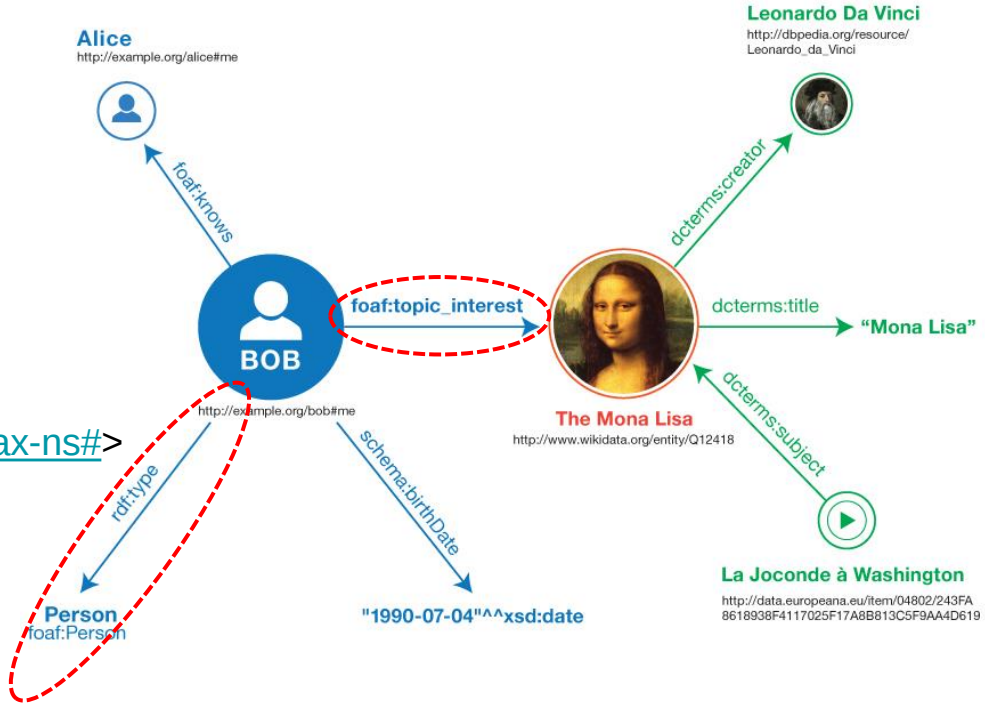
- ***RDF*** graph can be serialized using ***Turtle*** (Terse ***RDF*** Triple Language) syntax

```
@BASE   <http://example.org/>
@PREFIX foaf: <http://xmlns.com/foaf/0.1/>
@PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
@PREFIX schema: <http://schema.org/>
@PREFIX dcterms: <http://purl.org/dc/terms/>
@PREFIX wd: <http://www.wikidata.org/entity/>

<bob#me>
        a foaf:Person ;
        foaf:knows <alice#me> ;
        schema:birthDate "1990-07-04"^^xsd:date ;
        foaf:topic_interest wd:Q12418 .

wd:Q12418   S
        dcterms:title "Mona Lisa" ;
P       dcterms:creator <http://dbpedia.org/resource/Leonardo_da_Vinci> .   O

<http://data.europeana.eu/item/04802/243FA8618938F4117025F17A8B813C5F9AA4D619>
        dcterms:subject wd:Q12418 .
```



RDF Primer

- Use ***SPARQL*** (Protocol and ***RDF*** Query Language) to query ***RDF*** graph e.g. *"people interested in the Mona Lisa"*
- Query for multiple triple patterns



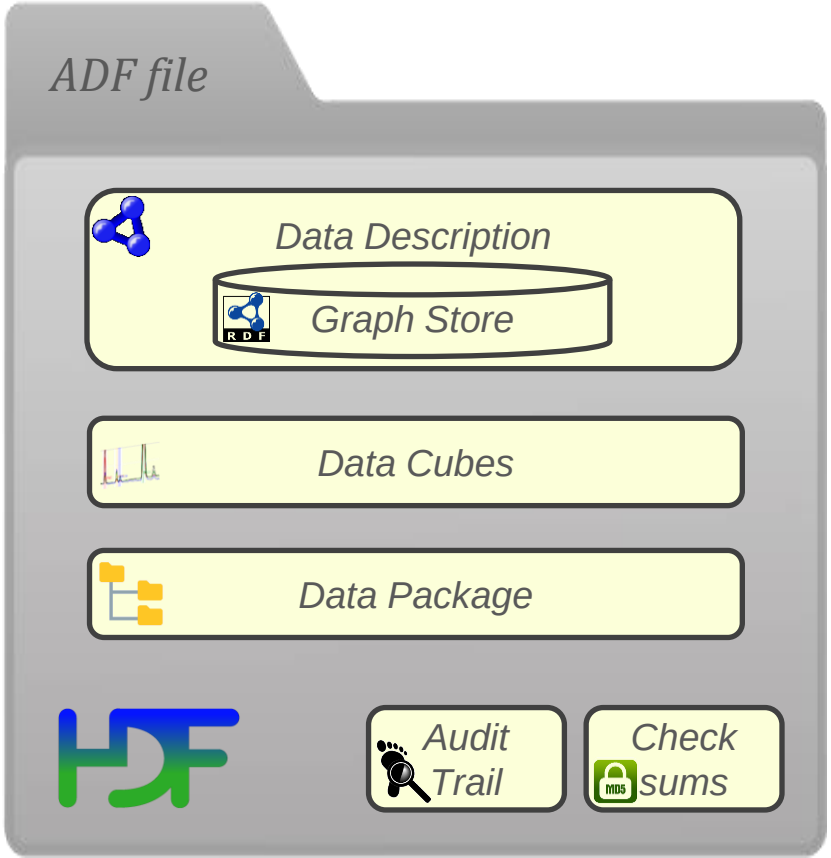

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX foaf:  <http://xmlns.com/foaf/0.1/>
PREFIX wd: <http://www.wikidata.org/entity>

SELECT ? name
WHERE {
   ?name rdf: type foaf: Person.
   ?name foaf: topic_interest wd:Q12418.
}
```

ADF file

ADF File Architecture and APIs

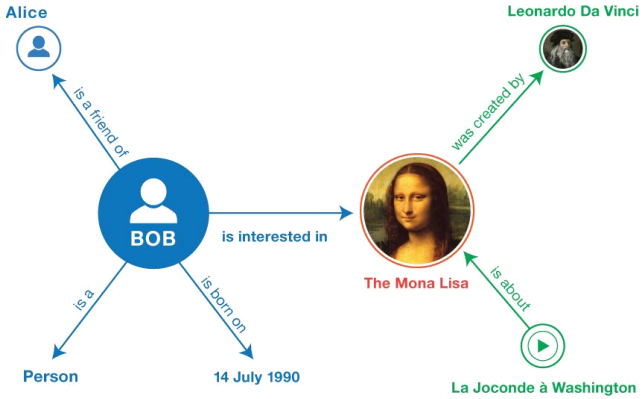
Generic ADF File

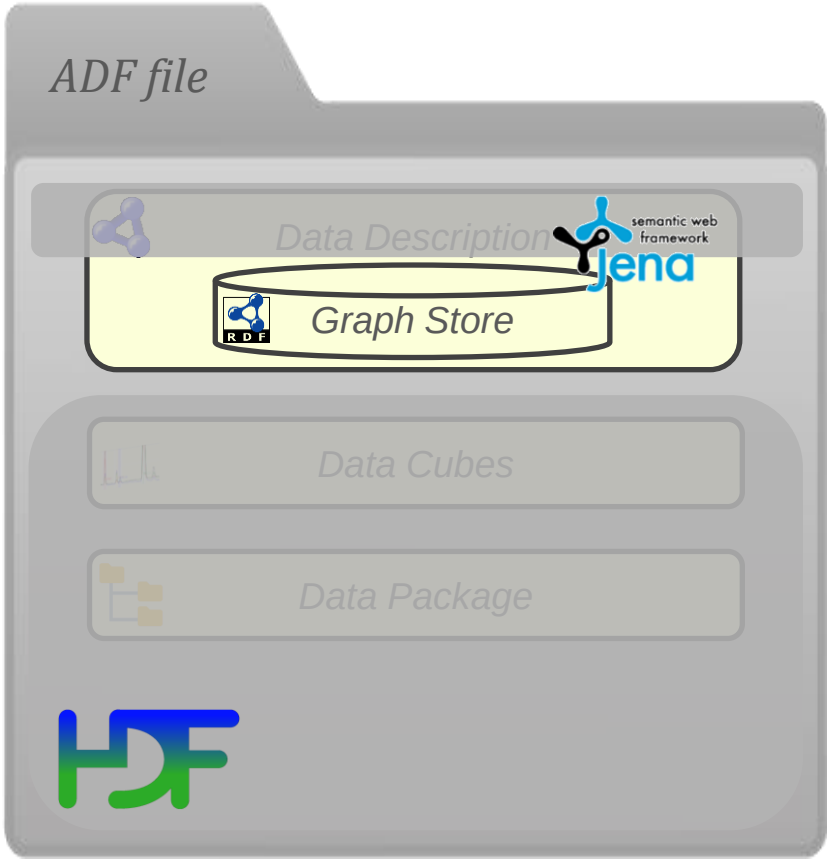
- ADF (Allotrope Data Format) File contains several components:
  - Data Description
  - Data Cubes
  - Data Package
  - Graph Store
  - Audit Trail (Audit/compliance features)
  - Checksums (Audit/compliance features)
- ADF File uses HDF5 as the underlying file format to store scientific data in a performant, consistent, and self-describing way



Generic ADF File

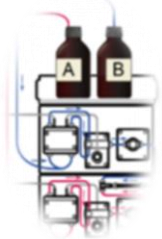
- Graph Store
  - Stores user defined graphs
  - Stores it in a form of ***RDF*** graph
  - Based on Apache Jena
- ***RDF*** Graph



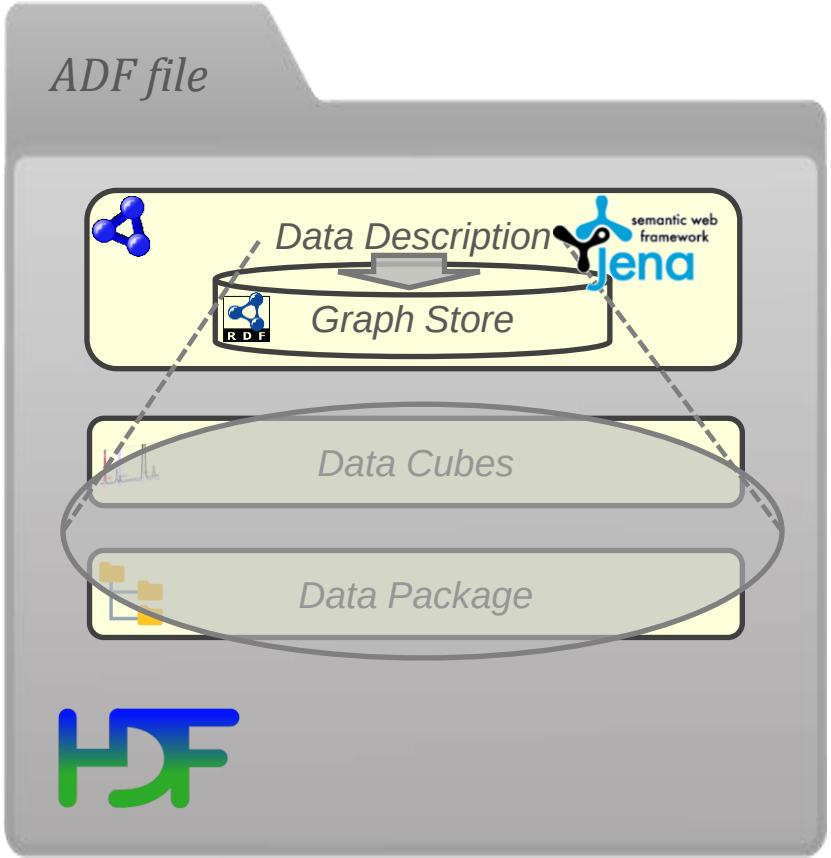


Generic ADF File

- Graph Store
  - Stores user defined graphs
  - Stores it in a form of ***RDF*** graph
  - Based on Apache Jena
- Data Description may contain
  - Contextual metadata about the Data Cube content
  - Contextual metadata about the Data Package content
  - Specifically for a Lab environment:
    - Results description
    - Sample description
    - Process description
    - Method description
    - Instrument description
    - Etc.

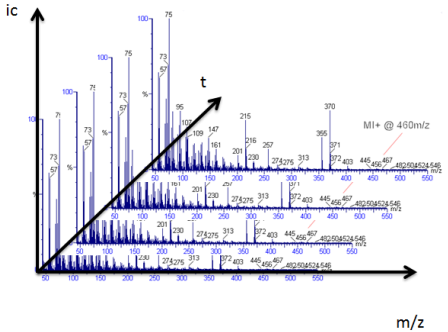


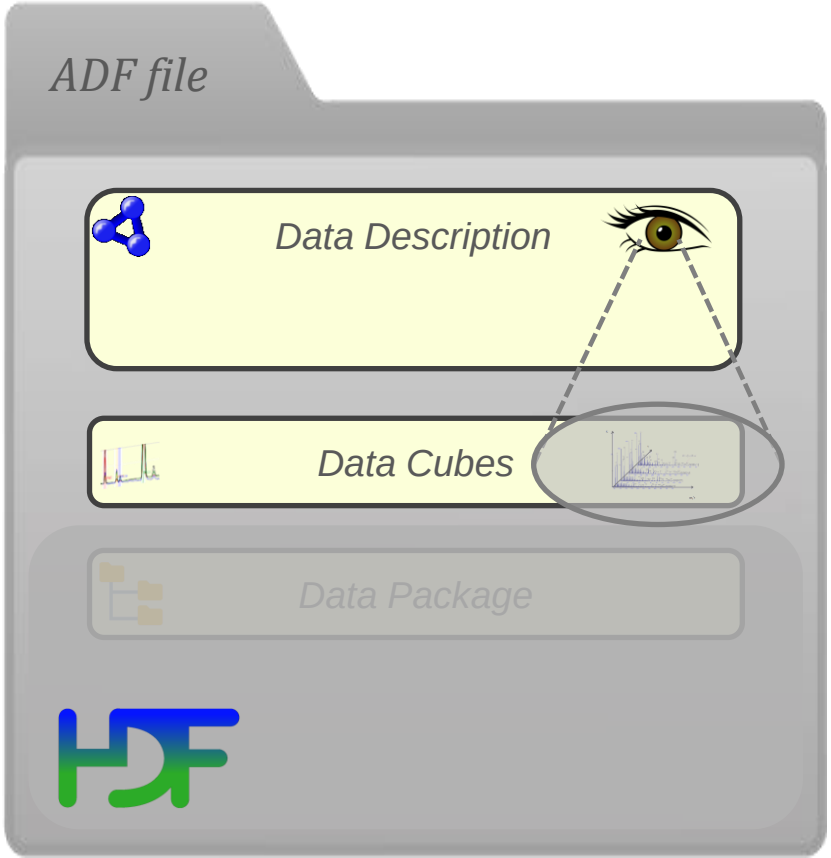
- The Data Description utilizes semantic web technology for a rich metadata and contextual information vs. HDF5 attributes



Generic ADF File

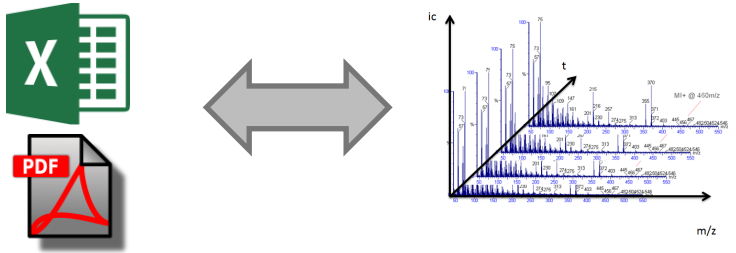
- Data Cubes
  - Contains multi-dimensional arrays of data
  - Specifically for a Lab environment it may contain measurements, results of an experiment or process data
  - Data Cube context (metadata) is described in the Data Description

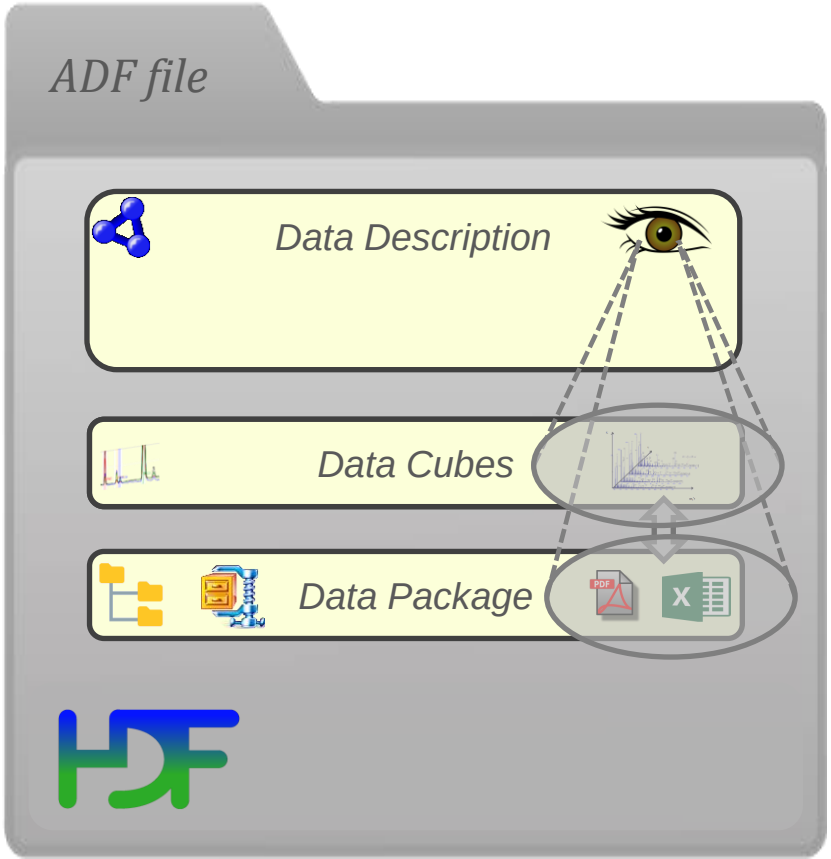




Generic ADF File

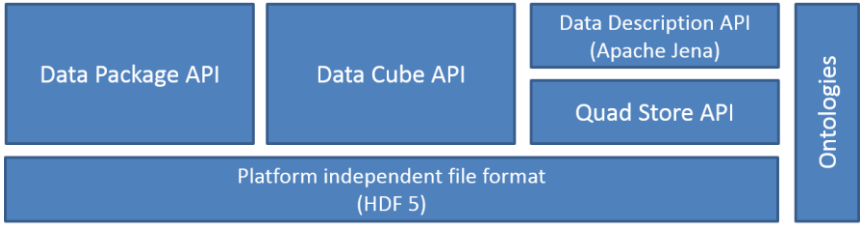
- Data Package
  - General purpose virtual file storage system
  - Can (optional) archive the same data in the Data Cubes but in its original native format and/or any other data
  - Data Package context (metadata) is described in the Data Description
  - Ensure data consistency and integrity of files during storage and transfer.

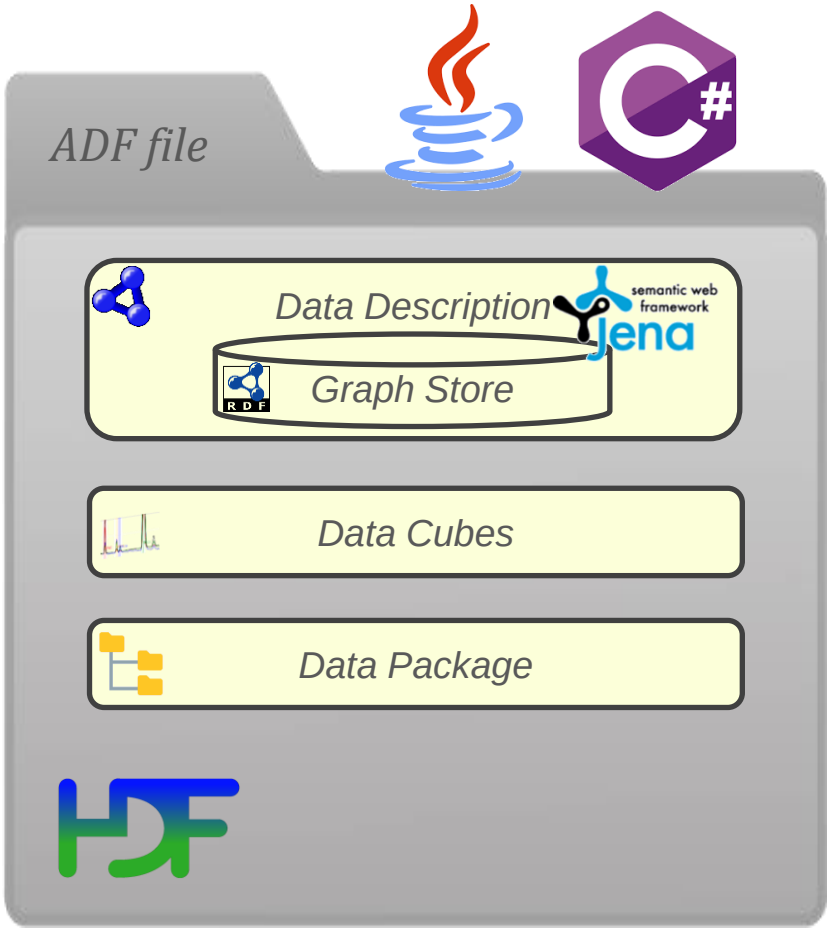




Generic ADF File

- ADF API Architecture

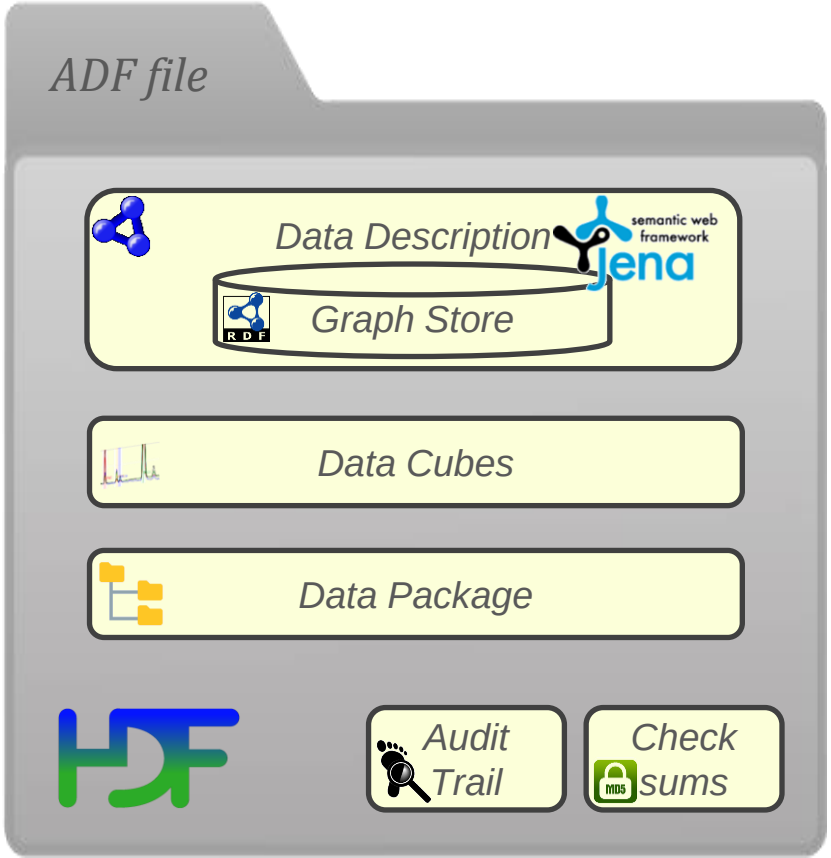




Generic ADF File

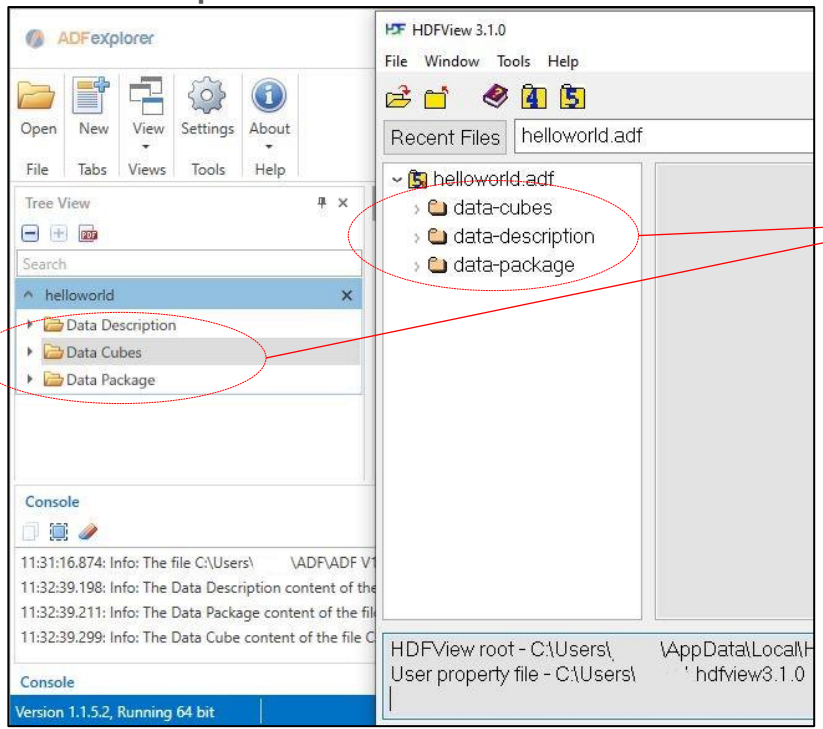
ADF internal structure utilizes HDF5 structure:

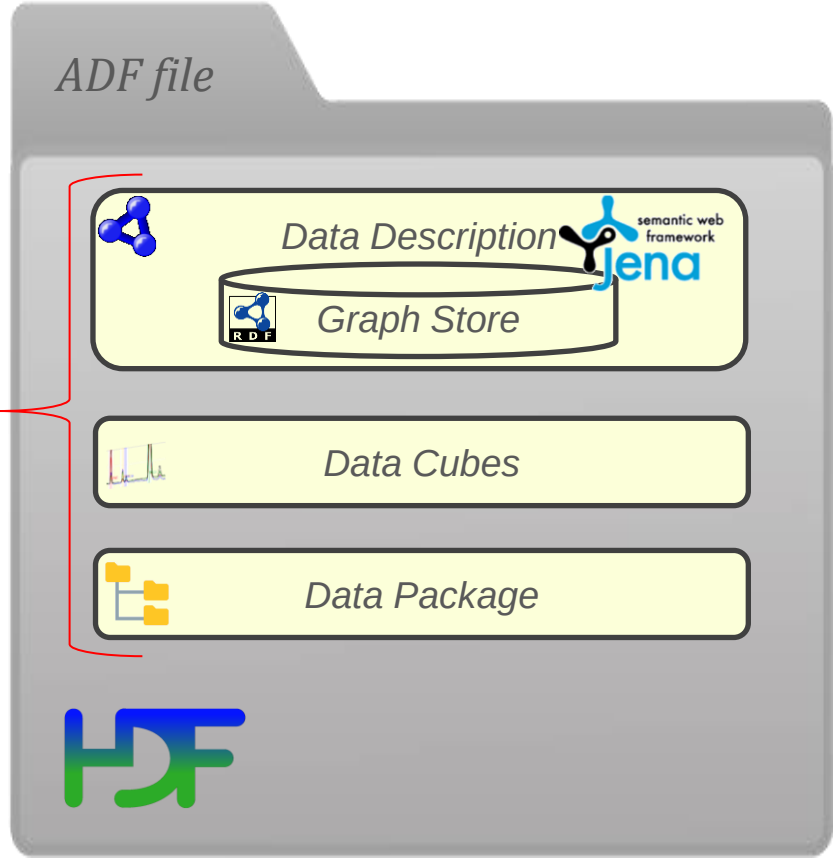
- Each of the three ADF layers is a top-level group in the HDF5 file
  - Data Description
  - Data Cubes
  - Data Description
- Auxiliary features are also stored as a top-level group in the HDF5 file
  - Named Graphs
  - Audit Trail



Generic ADF File

ADF Explorer vs HDF Viewer:

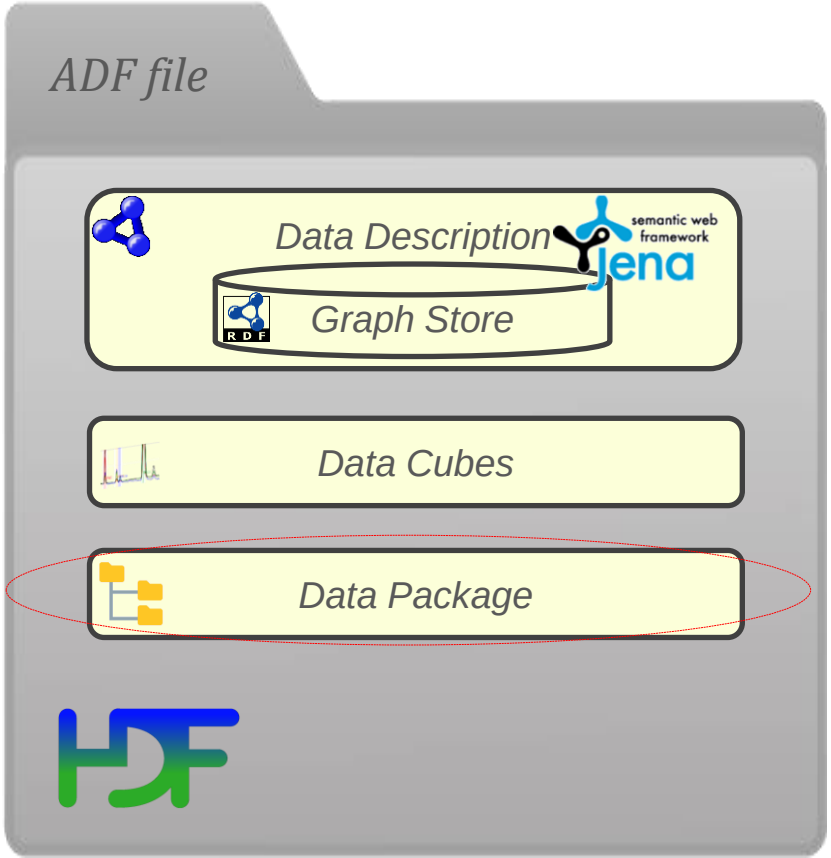




Generic ADF File

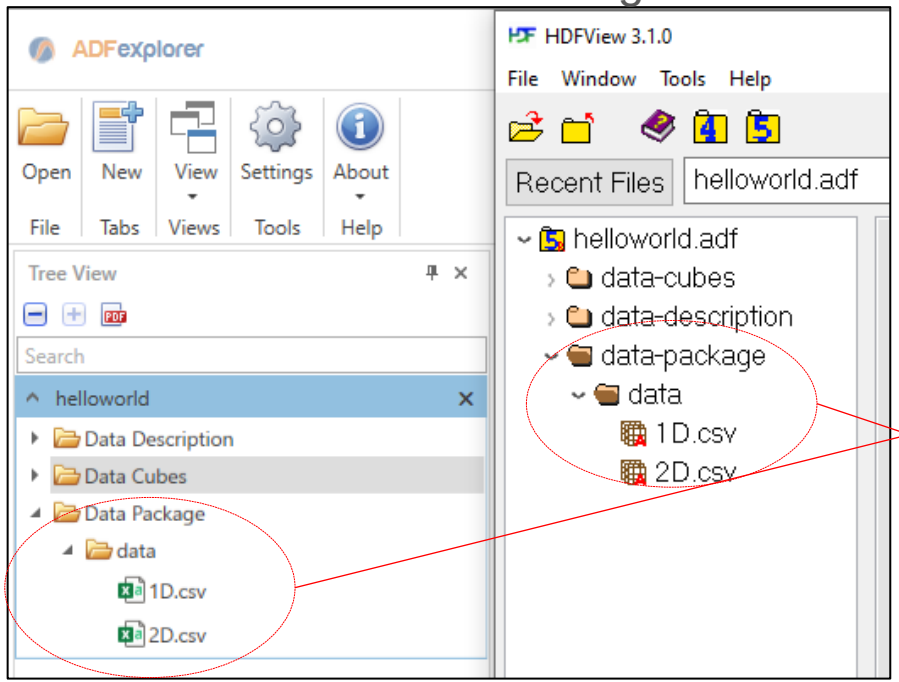
ADF structure – Data Package:

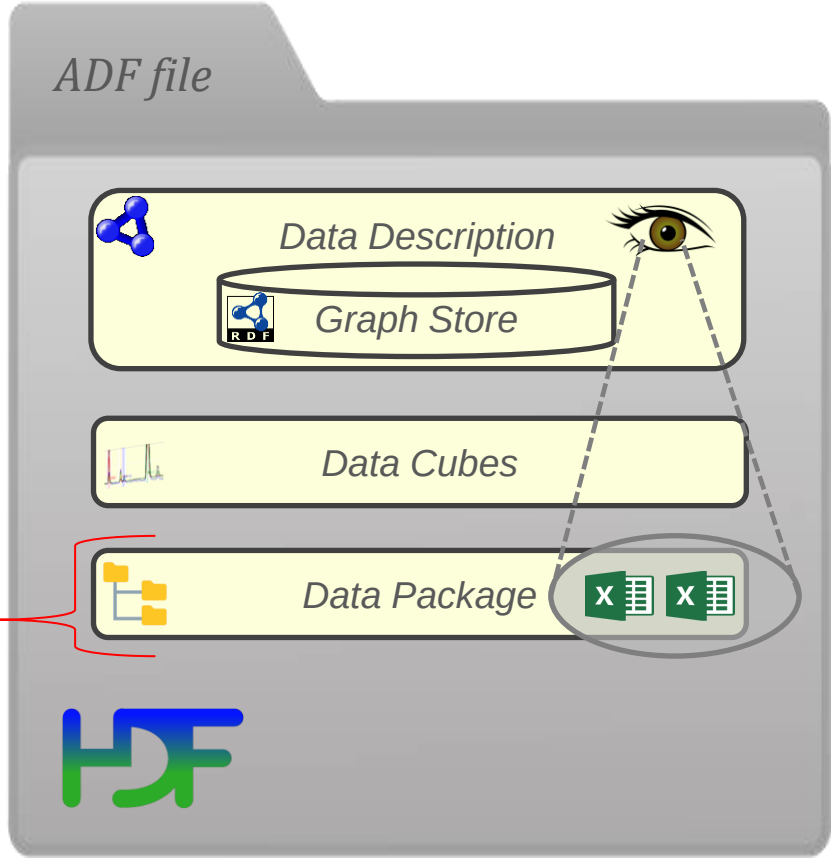
- ADF uses a top-level data package group to hold the file system layer
- Underneath, folders are represented as HDF Groups and files as HDF datasets
- Key metadata about individual files and folders are stored in the Data Description **automatically** by the ADF API



Generic ADF File

ADF structure – Data Package:

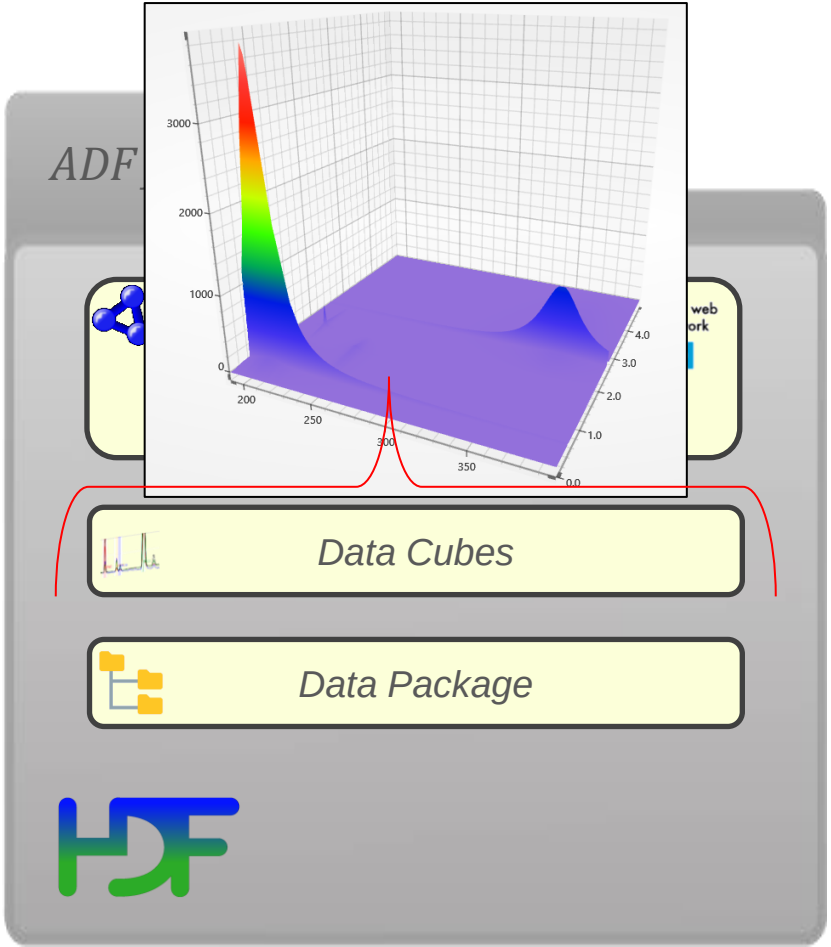




Generic ADF File

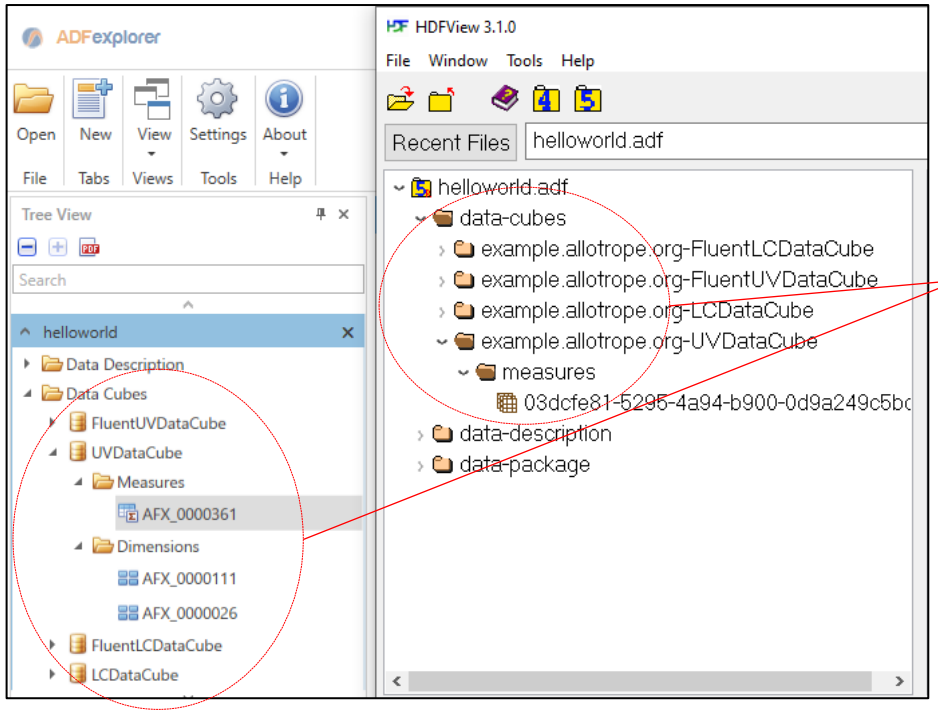
ADF structure – Data Cubes:

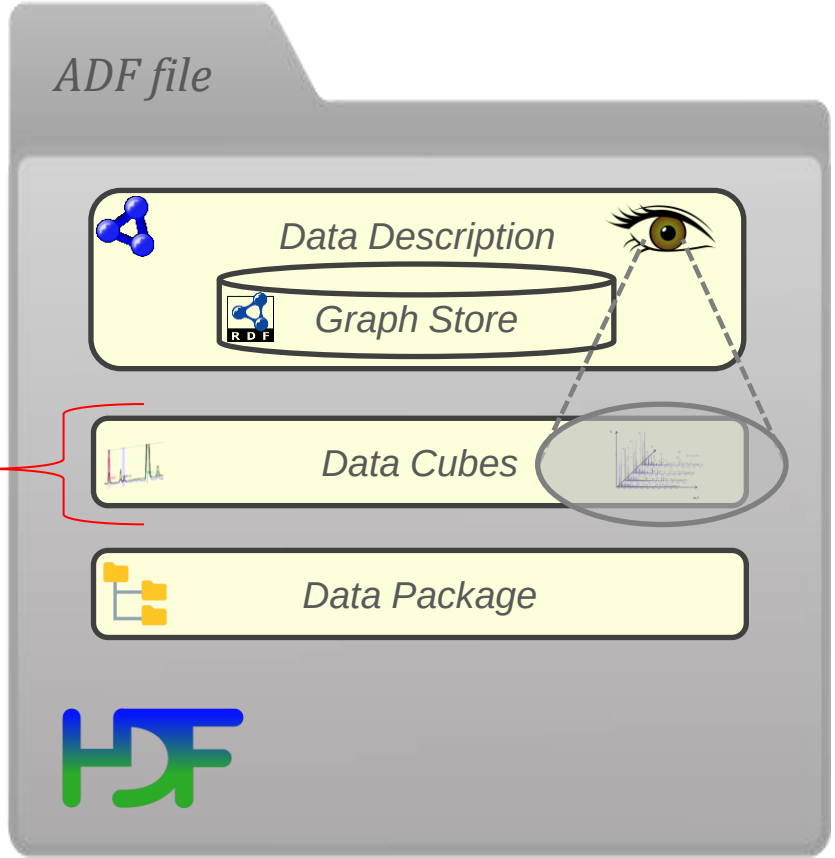
- RDF Data Cubes consist of measures, dimensions, and scales
  - *Measures* are the observed values for a given dataset, e.g. intensity or m/z
  - Dimensions are the controlled variables e.g. time or wavelength
  - Scales are the values of the dimensions, which may be linear, nonlinear, or entirely arbitrary
- All three are referenced in the Data Description such that metadata can be linked to them



Generic ADF File

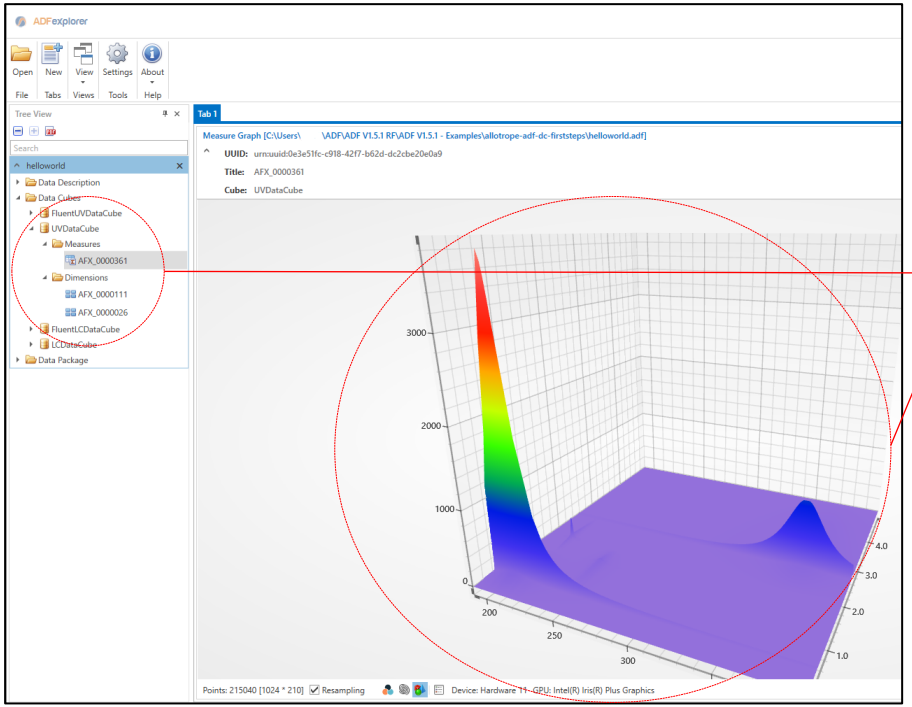
ADF structure – Data Cubes:

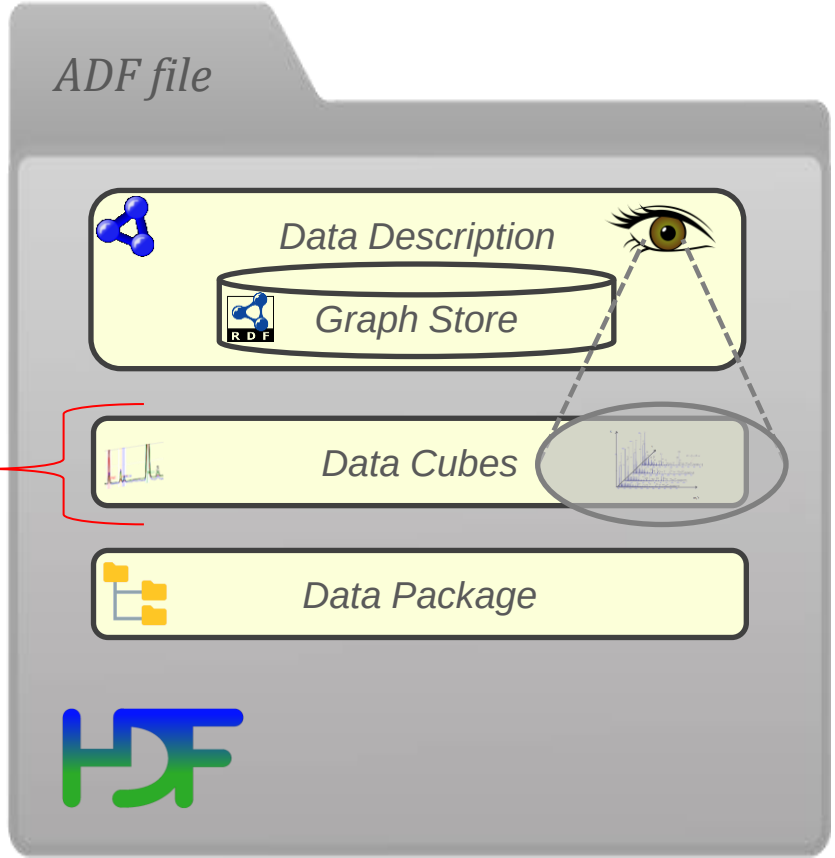




Generic ADF File

ADF structure – Data Cubes:

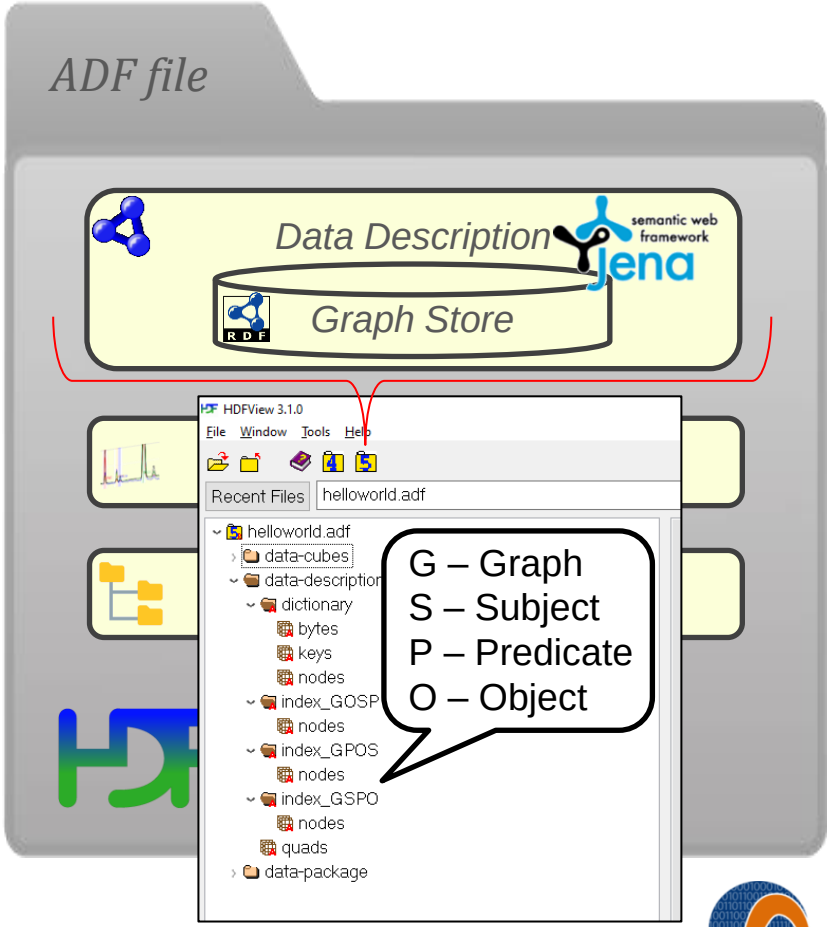




Generic ADF File

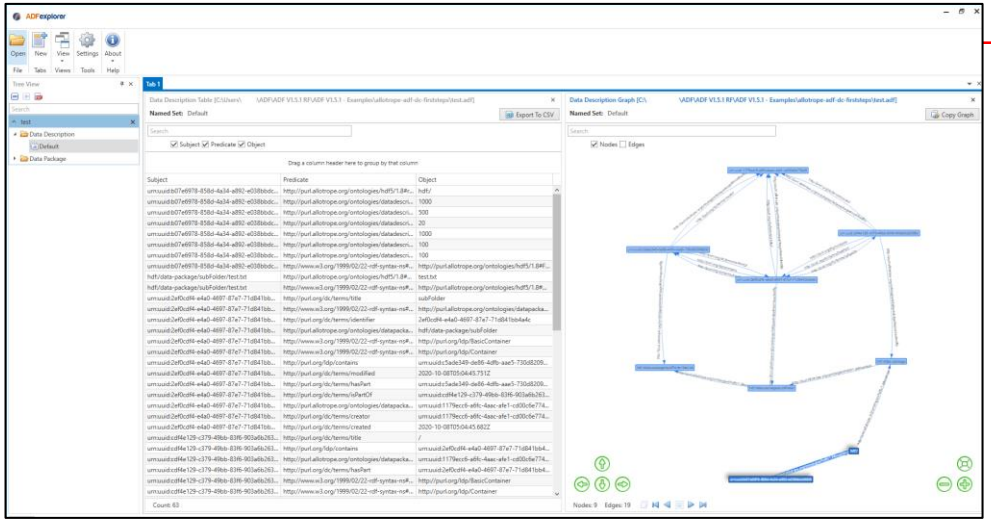
ADF structure – Data Description:

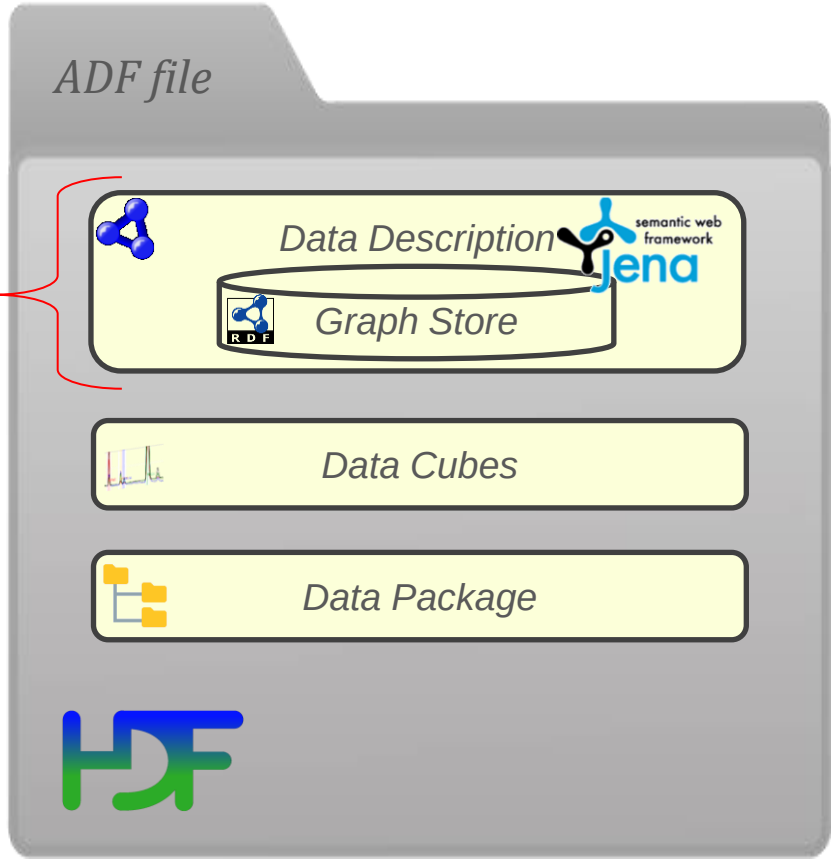
- HDF5 does not yet have native graph storage
- ADF emulates this through a combination of:
  - A dictionary
  - A list of quads (actually quints, including a deletion flag)
  - Multiple index listings for fast search

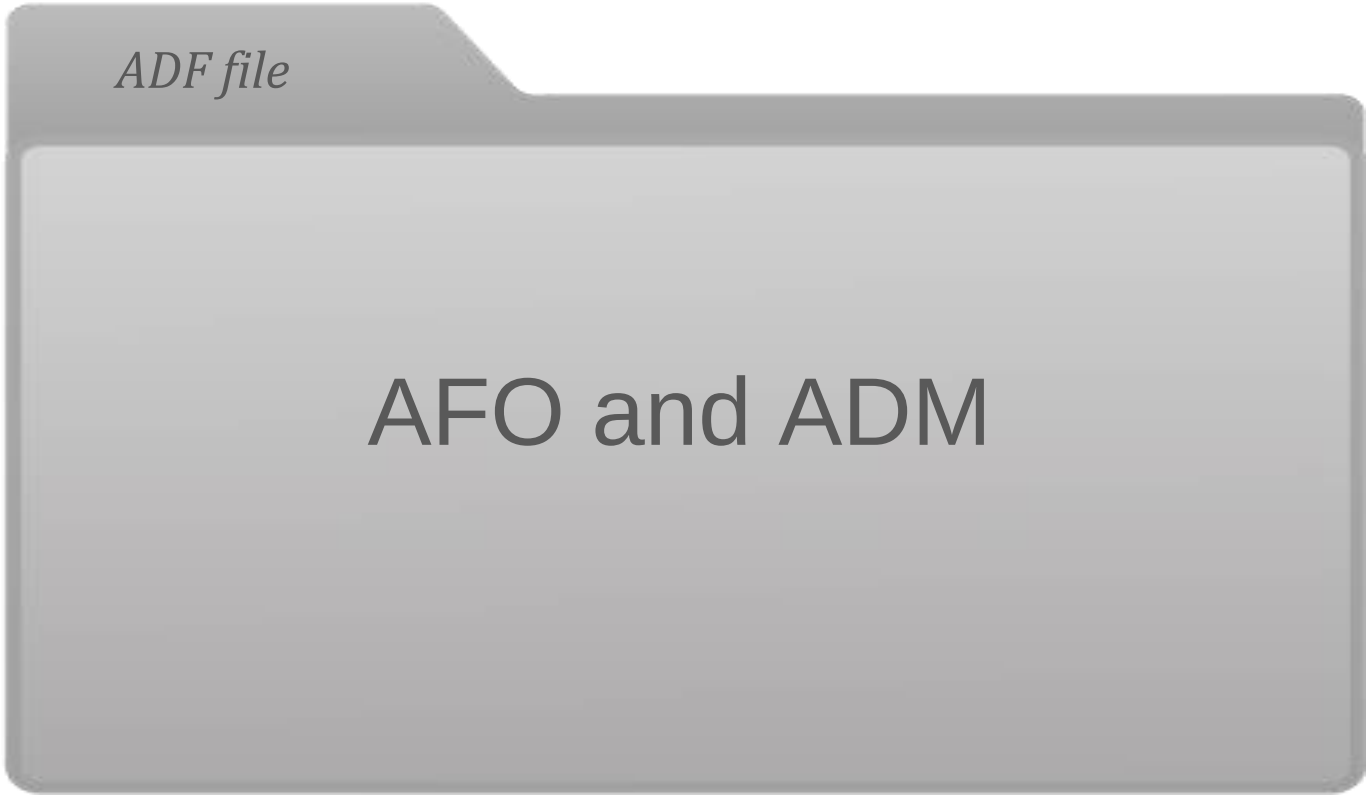


Generic ADF File

ADF structure – Data Description:

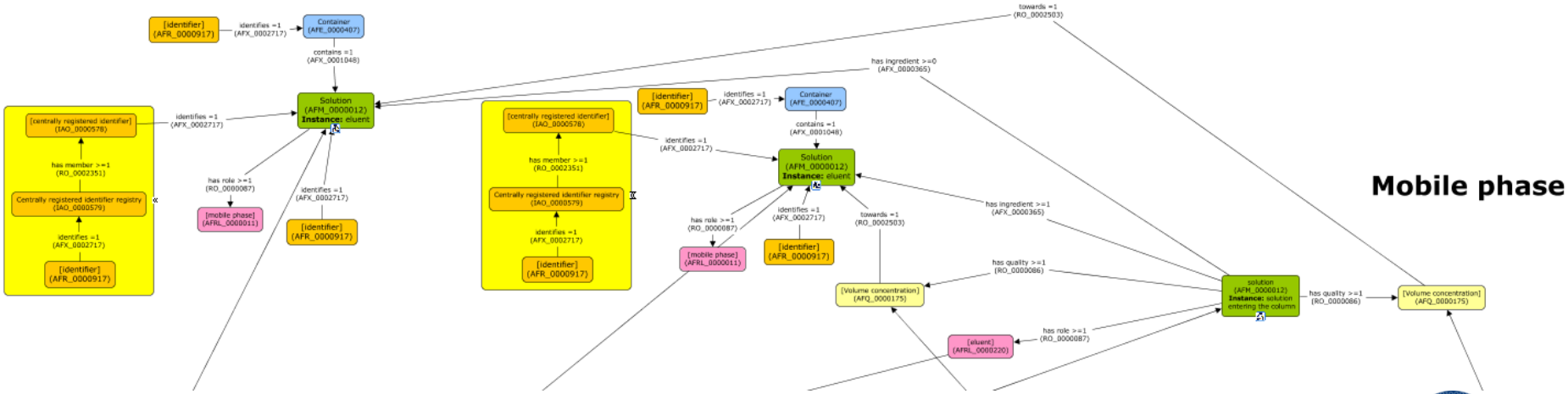






AFO: Allotrope Taxonomies and Ontologies

- To address the laboratory analytical processes:
  - The ***AFT*** - Allotrope Foundation Taxonomy formalizes the hierarchical relationships of terms
  - The ***AFO*** - Allotrope Foundation Ontology is an ontology suite that provides a standard vocabulary and semantic model representation



AFO: Terminology

- Entities within the vocabulary are uniquely identified with standard ***IRI****s*

device identifier

(http://purl.allotrope.org/ontologies/result#AFR_0002018)
synonyms: instrument id, machine id, device id, equipment id, equipment identifier, instrument identifier, machine identifier

A device identifier is an identifier that identifies some device. [Allotrope]

equipment serial number

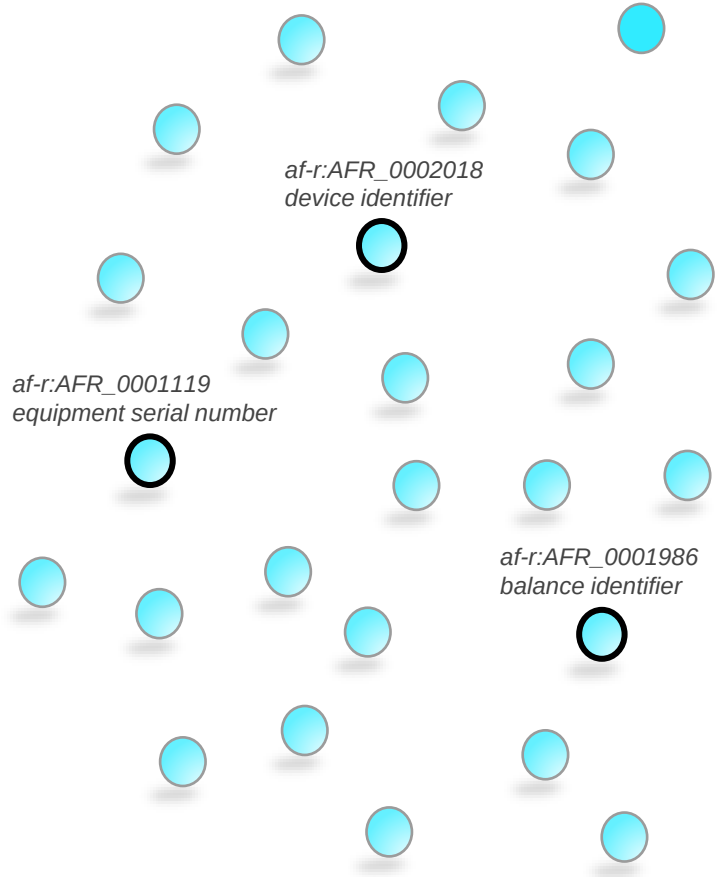
(http://purl.allotrope.org/ontologies/result#AFR_0001119)
synonyms: instrument serial number, device serial number, instrument id, machine id, machine serial number

Equipment serial number is measurement metadata that identifies an equipment used in the measuring by its serial number. [Allotrope]

balance identifier

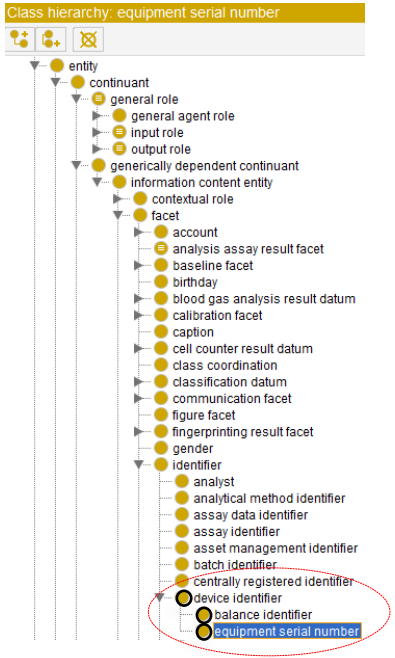
(http://purl.allotrope.org/ontologies/result#AFR_0001986)
synonyms: balance id

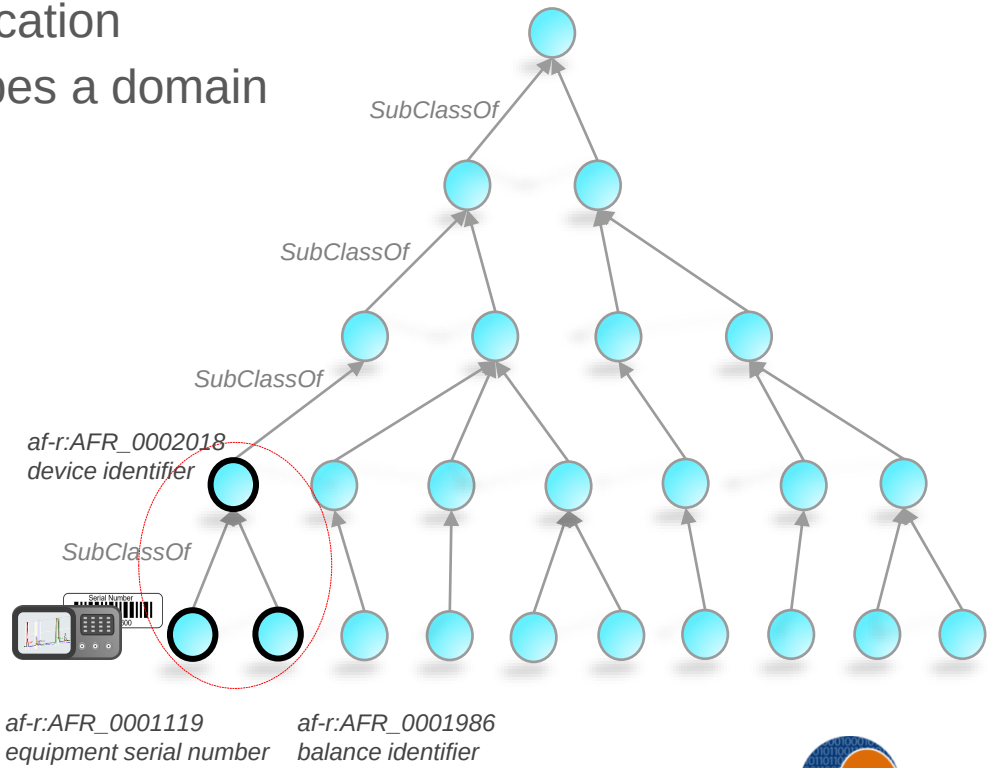
A balance identifier is a device identifier that identifies some balance. [Allotrope]



AFO: Terminology, Taxonomy

- Taxonomy is a hierarchical classification (organization) of terms that describes a domain





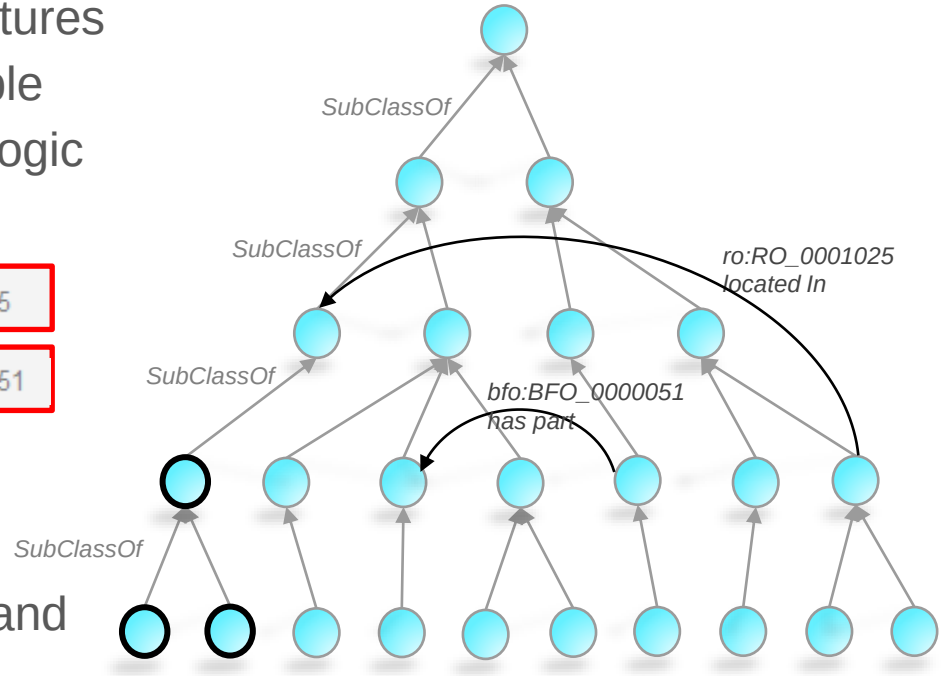
AFO: Terminology, Taxonomy and Ontology

- An ontology is a logic model that captures a domain in a machine understandable way including entities, relations and logic that controls relations.

located in — RO:0001025 — http://purl.obolibrary.org/obo/RO_0001025

has part — BFO:0000051 — http://purl.obolibrary.org/obo/BFO_0000051

- Ontologies provide an unconstrained vocabulary we can use to describe things (instances) in our open world and give them a meaning (= what it is)



AFO: Terminology, Taxonomy and Ontology

- From the graph we can learn that a specific equipment *"has contextual role in process"* (http://purl.allotrope.org/ontologies/property#AFX_0002731)

in process — http://purl.allotrope.org/ontologies/property#AFX_0002731

Not secure | purl.allotrope.org/ontologies/property#AFX_0002731

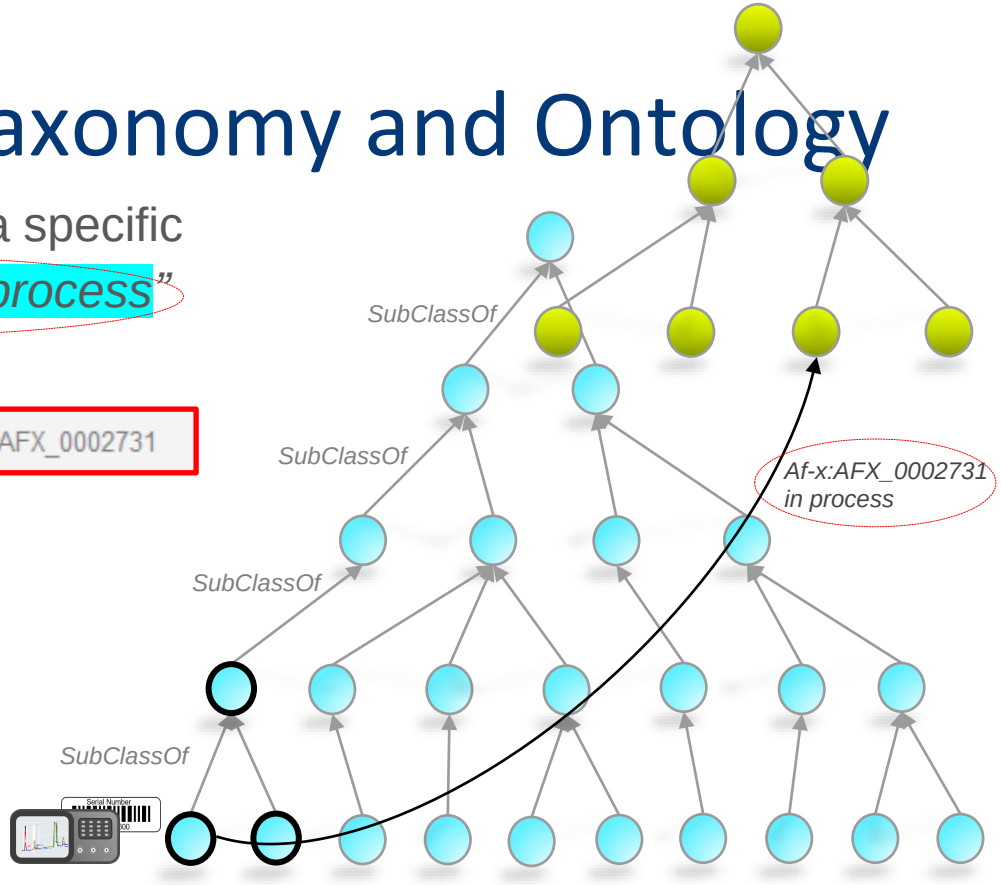
in process

(http://purl.allotrope.org/ontologies/property#AFX_0002731)

synonyms: has contextual role in process

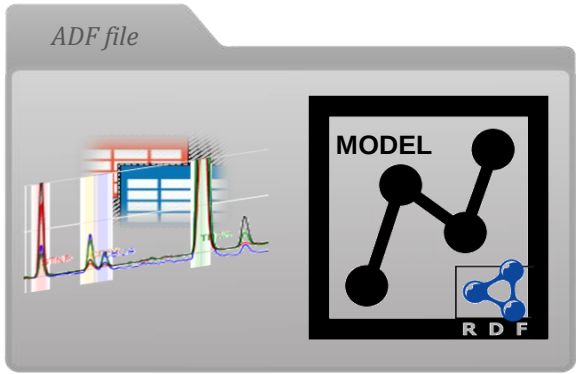
The relation between a conceptual role and the process where the contextual role is realized. [Allotrope]

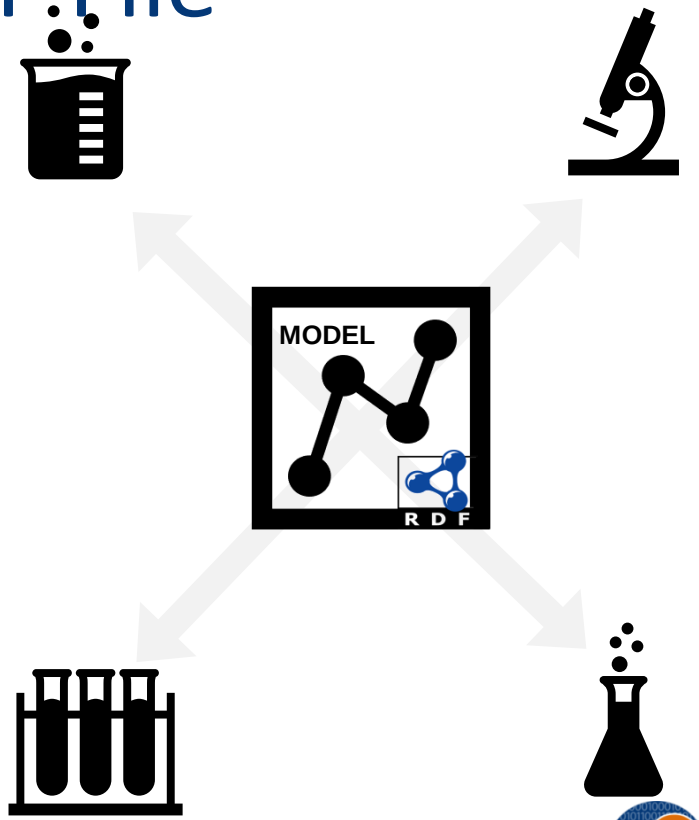
- Application can issue a query to *retrieve the list of equipment IDs that took part in a certain process*



ADM: Interoperable ADF File

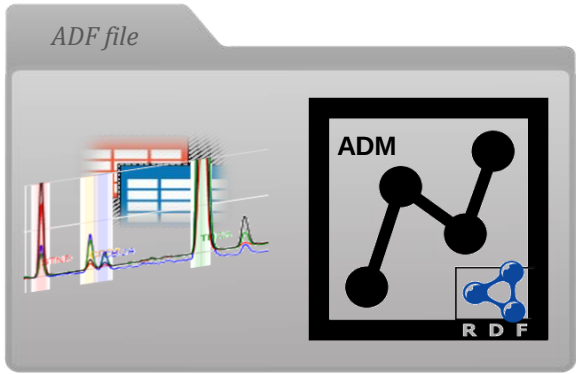
- A ***"Model"*** represents one or more "Use Case(s)"
- The ***"Model"*** uses ***RDF*** graph to describe the "Use Case(s)"

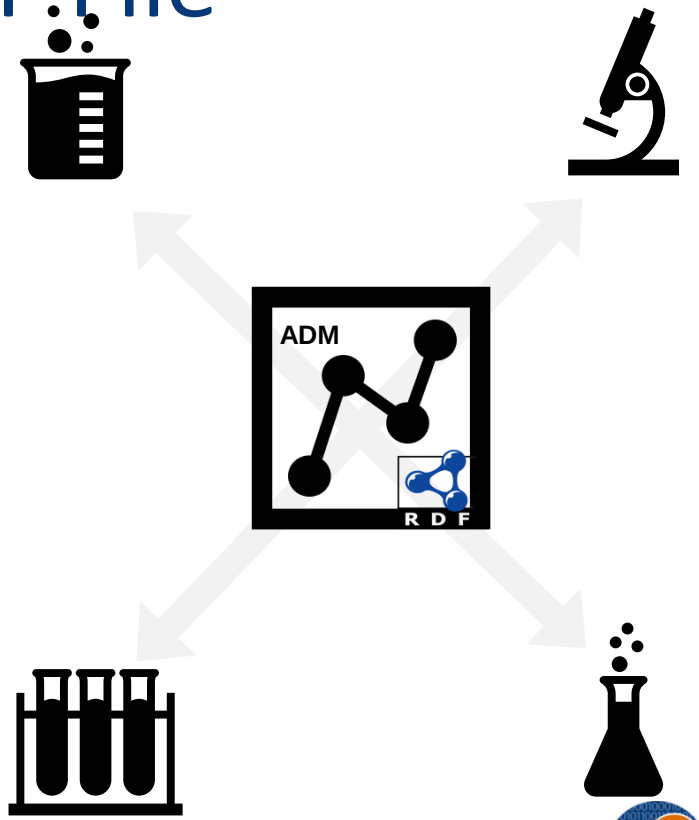




ADM: Interoperable ADF File

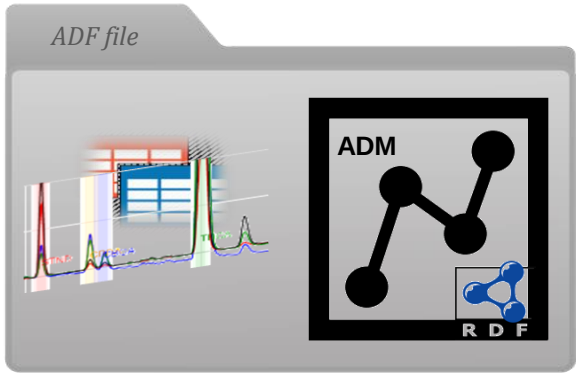
- An Allotrope standardized "***Model"*** for the pharmaceutical laboratory analytical processes is called ***ADM*** – Allotrope Data ***Model***

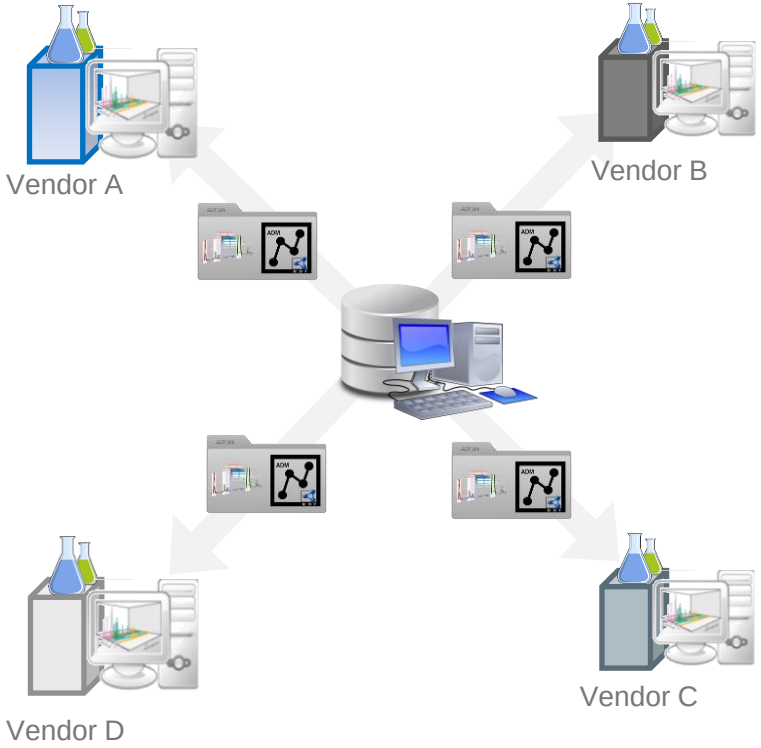




ADM: Interoperable ADF File

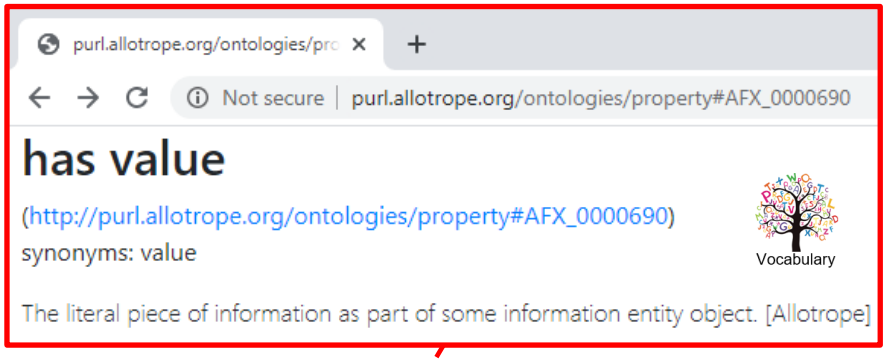
- Heterogenous vendor systems can seamlessly exchange ***ADF*** files (read and write) and process its ***Data*** that adheres to an associated standardized ***ADM***



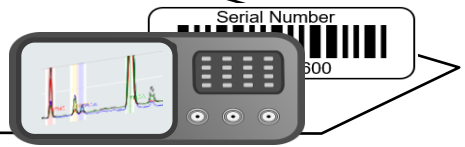


ADM: Shape File

- Contains a set of conditions
- Uses ***SHACL*** syntax:
  Shapes Constraint Language
- Example of a single condition (single shape) on the Equipment serial number:



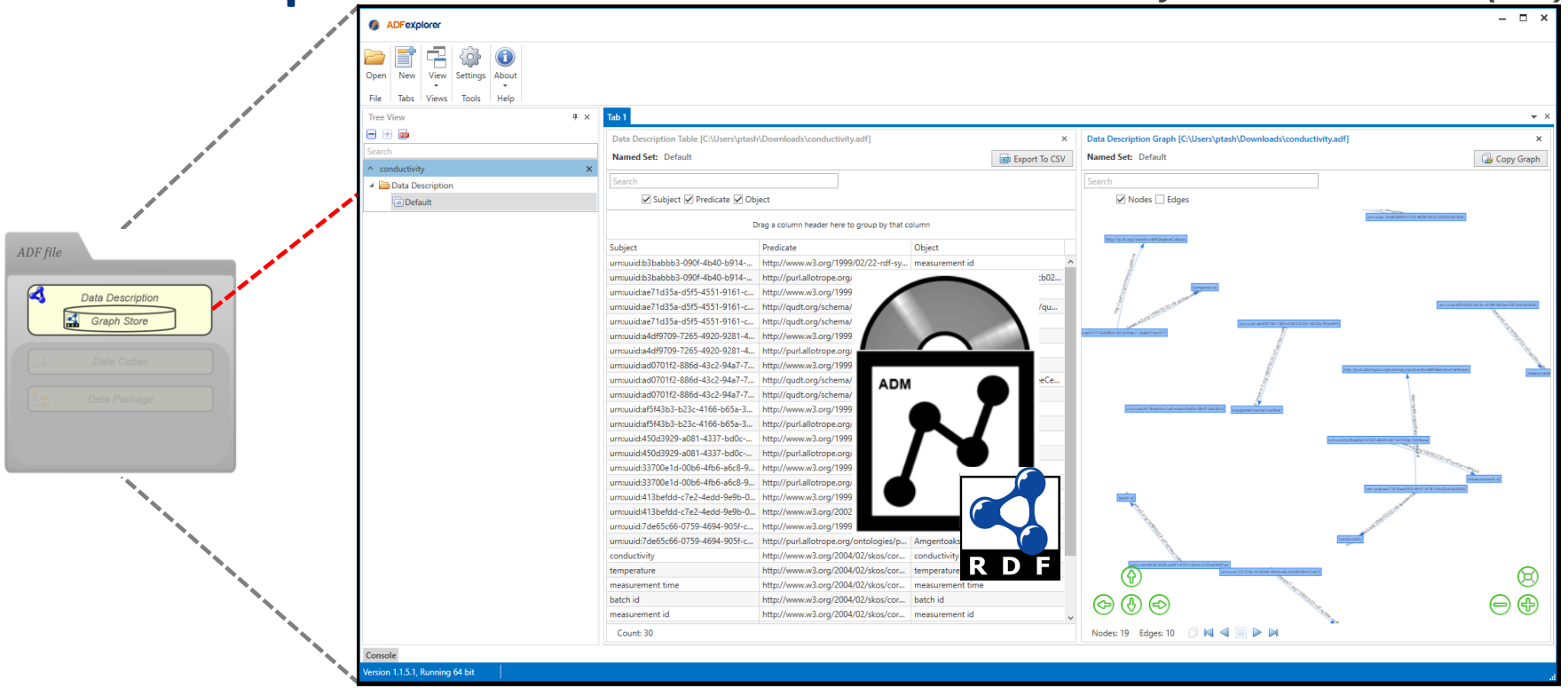

```
af-s:AFS_0000611
    rdf:type sh:PropertyShape ;
    sh:path <http://purl.allotrope.org/ontologies/property#AFX_0000690> ;
    sh:message "Equipment serial number MUST have exactly one value of type  xsd:string." ;
    sh:datatype xsd:string ;
    sh:maxCount 1 ;
    sh:minCount 1 ;
```



- Evaluate conformance of the ***ADM Data*** instance with the set of conditions in associated ***Shapes*** (SHACL)

Interoperable ADF File

Conductivity Measurement ADF (.adf)



ADF file

Thanks for your attention!

Allotrope Foundation Product Team

Amnon Ptashek amnon.ptashek@allotrope.org
Allotrope Foundation www.allotrope.org
email: more.info@allotrope.org