

Recent explorations for rapid data descriptor and ADF file development

Simplified SHACL Working Group

Ontology Working Group

Don Rolph, Amgen

Matt Potter-Racine, Amgen

Wes Schafer, MSD

Siping Wang, Tetrascience

Jon Hurley, BioVia

Ralph Hodgson, TopQuadrant

Objective

- Problem Statement
- Demonstrate Efficient Approach to creating ADF models for new instruments
- Examine application to multiple instruments
- Discussion of automation
- Forward looking efforts on next steps

Agenda:

- Don: Introduction to the problem: scalability of extending ADF including new terms and creation of SHACL data shapes
- Don cell counter and blood gas analyzer implementation
- Spin: demo of script for generating exemplar data descriptor and SHACL file from spreadsheet
- Don: summary/conclusions

Introduction to problem

- There are tens to hundreds of instrument types which are potential targets for leveraging the ADF framework
 - The approach used during 2017 to develop the ADF components for a new instrument class were quite time consuming and involved significant handcrafting by expert SMEs and expert Ontology specialists
 - We faced a scalability problem
- ❖Challenge: establish a procedure for creating the required ontology terms, an exemplar data descriptor, and a SHACL shape file in a period of two weeks or less

Some guiding principles

- A usable solution which we know how to execute is more useful than a better solution which we do not know how to execute
 - Do not let the perfect be the enemy of the good
- We will crawl before we walk, walk before we try to run
 - This drives an incremental evolutionary process in our efforts
 - We need to work to minimize technical debt (we can't eliminate it, but we need to minimize it consistent with making progress)
- There is an external reality of the data: our models must at a minimum be able to reflect this external reality of the data

Don: process for new instrument

- Establish draft of spreadsheet with proposed terms
- Have appropriate SMEs review for completeness
- Distribute for general review by as many parties as possible: e.g. ontology workstream extended team and as many SMEs and vendors as are participating
- Submit spreadsheet for ontology governance
- Create exemplar data descriptor and SHACL file
- Check with SPARQL query for roundtrip fidelity of the data

Spreadsheet to support Ontology Governance

Manufacturer Parameter	proposed prefLabel	Parameter Allotrope URI	definition
Measurement ID	measurement id	http://purl.allotrope.org/ontologies/result#AFR_0001121	Measurement time is measurement metadata about the name of the measurement.
RunDate	measurement time	http://purl.allotrope.org/ontologies/result#AFR_0000952	Measurement id is measurement metadata that identifies the measurement.
Operator	analyst	http://purl.allotrope.org/ontologies/result#AFR_001116	Analyst is measurement metadata about the name of the analyst.
Sample ID	sample id	http://purl.allotrope.org/ontologies/result#AFR_0001118	Sample id is measurement metadata that identifies the sample.
Machine ID	equipment serial number	http://purl.allotrope.org/ontologies/result#AFR_0001119	Equipment serial number is measurement metadata.
Batch ID	batch id	http://purl.allotrope.org/ontologies/result#AFR_0001120	Batch id is measurement metadata that identifies the batch.
Cell Type	cell type (cell counter)	http://purl.allotrope.org/ontologies/result#AFR_0001110	Cell type is cell counter parameter data that describes the cell type.
Dilution Ratio	cell density dilution factor	http://purl.allotrope.org/ontologies/result#AFR_0001109	Cell density dilution factor is a cell counter parameter data that describes the dilution ratio.
Viability	viability	http://purl.allotrope.org/ontologies/result#AFR_0001111	Viability is a cell counter result data that quantifies the viability.
Total Density	total cell density (cell counter)	http://purl.allotrope.org/ontologies/result#AFR_0001112	Total cell density is a cell counter result data that quantifies the total cell density.
Viable Density	viable cell density (cell counter)	http://purl.allotrope.org/ontologies/result#AFR_0001108	Viable cell density is a cell counter result data that quantifies the viable cell density.
Average Live Diameter	average live cell diameter (cell counter)	http://purl.allotrope.org/ontologies/result#AFR_0001113	Average live cell diameter is a cell counter result data that quantifies the average live cell diameter.
Total Live Count	viable cell count	http://purl.allotrope.org/ontologies/result#AFR_0001115	Viable cell count is a cell counter result data that quantifies the total live cell count.
Total Cell Count	total cell count	http://purl.allotrope.org/ontologies/result#AFR_0001114	Total cell count is a cell counter result data that quantifies the total cell count.
cells	cells	http://purl.allotrope.org/ontology/qudt-ext/unit#Cell	Count of cells in measurement.
10^6*cells/ml	10^6*cells/ml	http://purl.allotrope.org/ontology/qudt-ext/unit#MillionCellsPerL	Cells per unit volume



Microsoft Excel
Worksheet

Spreadsheet to create artifacts

Parameter Type	Parameter	Parameter PrefLabel	Parameter Allotrope URI	Required	Parameter Value	Type	Parameter Example Value	Parameter Unit	Syntax	Parameter Unit Qudt URI
metadata	Measurement ID	measurement id	http://purl.allotrope.org/ontologies/result#AFR_0001121	N	xsd:string	413befdd-c7e2-4edd-a0a6-0cf1-1b79ef				
metadata	RunDate	measurement time	http://purl.allotrope.org/ontologies/result#AFR_0000952	Y	xsd:dateTime	2015-09-24T03:47:13.001Z				
metadata	Operator	analyst	http://purl.allotrope.org/ontologies/result#AFR_0001116	N	xsd:string	Amgentoaks1				
metadata	Sample ID	sample id	http://purl.allotrope.org/ontologies/result#AFR_0001118	Y	xsd:string	unknown-10				
metadata	Machine ID	equipment serial number	http://purl.allotrope.org/ontologies/result#AFR_0001119	N	xsd:string	serial-number				
metadata	Batch ID	batch id	http://purl.allotrope.org/ontologies/result#AFR_0001120	N	xsd:string	batch-number				
Process parameter d: Cell Type	cell type (cell counter)		http://purl.allotrope.org/ontologies/result#AFR_0001110	N	xsd:string	CHO				
Process parameter d: Dilution Ratio	cell density dilution factor		http://purl.allotrope.org/ontologies/result#AFR_0001109	N	xsd:double	1				
Results Data	Viability	viability	http://purl.allotrope.org/ontologies/result#AFR_0001111	Y	xsd:double	0.1	percent	http://qudt.org/vocab/unit#Percent		
Results Data	Total Density	total cell density (cell counter)	http://purl.allotrope.org/ontologies/result#AFR_0001112	N	xsd:double	102.24	10^6 *cells/ml	http://purl.allotrope.org/ontology/qudt-ext/unit#MillionCellsPerML		
Results Data	Viable Density	viable cell density (cell counter)	http://purl.allotrope.org/ontologies/result#AFR_0001108	Y	xsd:double	0.05	10^6 *cells/ml	http://purl.allotrope.org/ontology/qudt-ext/unit#MillionCellsPerML		
Results Data	Average Live Diameter	average live cell diameter (cell count)	http://purl.allotrope.org/ontologies/result#AFR_0001113	N	xsd:double	21.07	micrometer	http://qudt.org/vocab/unit#Micrometer		
Results Data	Total Live Count	viable cell count	http://purl.allotrope.org/ontologies/result#AFR_0001115	N	xsd:double	1	cell	http://purl.allotrope.org/ontology/qudt-ext/unit#Cell		
Results Data	Total Cell Count	total cell count	http://purl.allotrope.org/ontologies/result#AFR_0001114	Y	xsd:double	1972	cell	http://purl.allotrope.org/ontology/qudt-ext/unit#Cell		



Microsoft Excel Worksheet



C:\Users\rolphd\Documents\Cell Count



C:\Users\rolphd\Documents\Cell Count



C:\Users\rolphd\Documents\Cell Count

Spreadsheet

Data Descriptor

SHACL File

ADF File

Extension to include proprietary fields

Data Type	Manufacturer Parameter	proposed preLabel	Parameter Allotrope URI	definition	entailment	Required	Example Parameter Value	Parameter Units of measure	Parameter Units of measure URI
metadata	Measurement ID	measurement id		Measurement time is measurement metadata about the date/time of the measurement. [Allotrope]	413befdd-c7e2-4edd-9e9b-06cf1cb0283f				
	RunDate	measurement time	http://purl.allotrope.org/ontologies/result#AFR_0000952	Measurement id is measurement metadata that identifies the measuring run. [Allotrope]	y	2015-09-24T03:47:13.0Z		xsd:dateTimeStamp	
Operator	analyst			Analyst is measurement metadata about the name or identifier of a person that has the role of an analyst in the measurement. [Allotrope]				xsd:string	
Sample ID	sample id			Sample id is measurement metadata that identifies a sample being measured. [Allotrope]			unknown-10	xsd:string	
Machine ID	equipment serial number			Equipment serial number is measurement metadata that identifies an equipment used in the measurement by its serial-number. [Allotrope]				xsd:string	
Batch ID	batch id			Batch id is measurement metadata that identifies the batch where a sample is taken from for being measured. [Allotrope]				xsd:string	
Process parameter data	Cell Type	cell type (cell counter)	http://purl.allotrope.org/ontologies/result#AFR_0000110	Cell type is cell counter parameter data that describes AFX_0000399 ("has input")			CHO		
	Dilution Ratio	cell density dilution factor	http://purl.allotrope.org/ontologies/result#AFR_0000109	Cell density dilution factor is a cell counter parameter data that describes AFX_0000399 ("has input")			1	xsd:double	
Results Data	Viability	viability	http://purl.allotrope.org/ontologies/result#AFR_0000111	Viability is a cell counter result data that quantifies AFX_0000395 ("has output")	y	0.1 percent		http://qudt.org/vocab/unit#Percent	
	Total Density	total cell density (cell counter)	http://purl.allotrope.org/ontologies/result#AFR_0000108	Total cell density is a cell counter result data that quantifies AFX_0000395 ("has output")			102.4 10^6 cells/ml	http://purl.allotrope.org/ontology/qudt-ext/unit#MillionCellsPerMilliliter	
	Visible Density	visible cell density (cell counter)	http://purl.allotrope.org/ontologies/result#AFR_0000109	Visible cell density is a cell counter result data that quantifies AFX_0000395 ("has output")	y	0.05 10^6 cells/ml		http://purl.allotrope.org/ontology/qudt-ext/unit#MillionCellsPerMilliliter	
	Average Live Diameter	average live cell diameter (cell counter)	http://purl.allotrope.org/ontologies/result#AFR_0000113	Average live cell diameter is a cell counter result data that quantifies AFX_0000395 ("has output")			21.07 micrometer	http://qudt.org/vocab/unit#Micrometer	
	Total Live Count	visible cell count	http://purl.allotrope.org/ontologies/result#AFR_0000115	Visible cell count is a cell counter result data that quantifies AFX_0000395 ("has output")			1 cell	http://purl.allotrope.org/ontology/qudt-ext/unit#Cell	
	Total Cell Count	total cell count	http://purl.allotrope.org/ontologies/result#AFR_0000114	Total cell count is a cell counter result data that quantifies AFX_0000395 ("has output")	y		1972 cell	http://purl.allotrope.org/ontology/qudt-ext/unit#Cell	
Vendor Specific terms									
Process parameter data	Cell Inspection Type	cell counter method	http://example.org/novacd#NOVACVD_00000000001	method used during test	y	method-1		xsd:string	
	Flow Time	event start time	http://example.org/novacd#NOVACVD_00000000002	duration of time required to capture all images used in analysis	y	392 s		http://qudt.org/vocab/unit#SecondTime	
	Mixing Routine	mixing routine	http://example.org/novacd#NOVACVD_00000000003	routine used during mixing process	y	mixing-routine-1		xsd:string	
	Number of Images	number of images	http://example.org/novacd#NOVACVD_00000000004	number of images captured during process	y	37		xsd:integer	
	Pre-dilution multiplier	pre-dilution multiplier	http://example.org/novacd#NOVACVD_00000000005	multiplier on results prior to dilution of sample	y	1		xsd:double	
	Tray Location	tray location	http://example.org/novacd#NOVACVD_00000000006	location of sample in sample tray used during measurement	y	tray-slot-7		xsd:string	
	Vessel ID	vessel id	http://example.org/novacd#NOVACVD_00000000007	Vessel ID of sample used during measurement	y	vessel-id1		xsd:string	
Results Data	Live Cell Count Standard Deviation	live cell count standard deviation	http://example.org/novacd#NOVACVD_00000000008	standard deviation of live cell counts	y	0.03 cell		http://purl.allotrope.org/ontology/qudt-ext/unit#Cell	
	Live Standard Deviation	live cell diameter standard deviation	http://example.org/novacd#NOVACVD_00000000009	standard deviation of live cell diameter	y	0.13 micrometer		http://qudt.org/vocab/unit#Micrometer	
	Total Density	vendor total density	http://example.org/novacd#NOVACVD_00000000010	vendor supplied value of total density	y	1022.4 10^6 cells/ml		http://operations.refdata.amgen.com/id/units-of-measure#AMUOM_000000000023	
	Viable Density	vendor viable density	http://example.org/novacd#NOVACVD_00000000011	vendor supplied value of viable density	y	0.5 10^6 cells/ml		http://operations.refdata.amgen.com/id/units-of-measure#AMUOM_000000000023	



Microsoft Excel
Worksheet

BGA Model: Governance and Artifacts completed in 7 days

Data Type	Manufacturer Parameter	proposed preLabel	Parameter Allotrope URI	Definition	entailment	Description: for clarity these should be thought of as data fields coming from an instrument. Their context requires either the ontology or an extended data descriptor to clearly define the terms	Required	Example Parameter Value	Parameter Units of measure	Parameter Units of measure URI
metadata	Measurement ID	measurement id	http://purl.allotrope.org/ontologies/resultNAFR_0001121	Measurement time is measurement metadata about the date/time of the measurement. [Allotrope]		unique id number for ADF file		413befdd-c7e2-4edd-9e9b-06cf1cb0283f		
	RunDate	measurement time	http://purl.allotrope.org/ontologies/resultNAFR_0000952	Measurement id is measurement metadata that identifies the measuring run. [Allotrope]		official date associated with measurement	y	2015-09-24T03:47:13.02		xsd:dateTimeStamp
Operator	analyst		http://purl.allotrope.org/ontologies/resultNAFR_0001116	Analyst is measurement metadata about the name or identifier of a person that has the role of an analy		person who is responsible for measurement		Analystoak1		xsd:string
Sample ID	sample id		http://purl.allotrope.org/ontologies/resultNAFR_0001118	Sample id is measurement metadata that identifies a sample being measured. [Allotrope]		id for sample tested. Is not necessarily unique	y	unknown-10		xsd:string
	Machine ID	equipment serial number	http://purl.allotrope.org/ontologies/resultNAFR_0001119	Equipment serial number is measurement metadata that identifies an equipment used in the measur		id number for equipment used in testing to support traceability		serial-number		xsd:string
	Batch ID	batch id	http://purl.allotrope.org/ontologies/resultNAFR_0001120	Batch id is measurement metadata that identifies the batch where a sample is taken from for being me		manufactured material		batch-number		xsd:string
Results Data	Partial Pressure CO2	pCO2	http://purl.allotrope.org/ontologies/resultNAFR_0001140	PCO2 is a bio assay result data that states the partial afx:AFX_0000395 ("has output")		partial pressure of CO2 measured in sample	y	mmHG		http://qudt.org/vocab/unit#MillimeterOfMercury
	Partial Pressure O2	pO2	http://purl.allotrope.org/ontologies/resultNAFR_0001141	P02 is a bio assay result data that states the partial afx:AFX_0000395 ("has output")		partial pressure of O2 measured in sample	y	mmHG		http://qudt.org/vocab/unit#MillimeterOfMercury
	pH	pH	http://purl.allotrope.org/ontologies/resultNAFR_0001142	PH is a bio assay result data that quantifies the acid afx:AFX_0000395 ("has output")		pH measured in sample	y			xsd:integer



Microsoft Excel Worksheet



Microsoft Excel Worksheet



C:\Users\rolphd\ments\BloodGasAr



C:\Users\rolphd\ments\BloodGasAr



C:\Users\rolphd\ments\BloodGasAr

Governance
Spreadsheet

Artifacts
Worksheet

Data Descriptor

SHACL File

ADF file

Strawman for Balance

Data Type	Manufacturer Parameter	proposed prefLabel	Parameter Allotrope URI	definition
metadata	Measurement ID	measurement id	http://purl.allotrope.org/ontologies/result#AFR_0001121	Measurement time is measurement metadata about the name of the measurement.
	RunDate	measurement time	http://purl.allotrope.org/ontologies/result#AFR_0000952	Measurement id is measurement metadata that identifies the measurement.
	Operator	analyst	http://purl.allotrope.org/ontologies/result#AFR_0001116	Analyst is measurement metadata about the name of the analyst.
	Sample ID	sample id	http://purl.allotrope.org/ontologies/result#AFR_0001118	Sample id is measurement metadata that identifies the sample.
	Machine ID	equipment serial number	http://purl.allotrope.org/ontologies/result#AFR_0001119	Equipment serial number is measurement metadata that identifies the machine.
	Batch ID	batch id	http://purl.allotrope.org/ontologies/result#AFR_0001120	Batch id is measurement metadata that identifies the batch.
Process parameter data				
Results Data		gross weight		
		tare weight		
		net weight		
		stable weight		

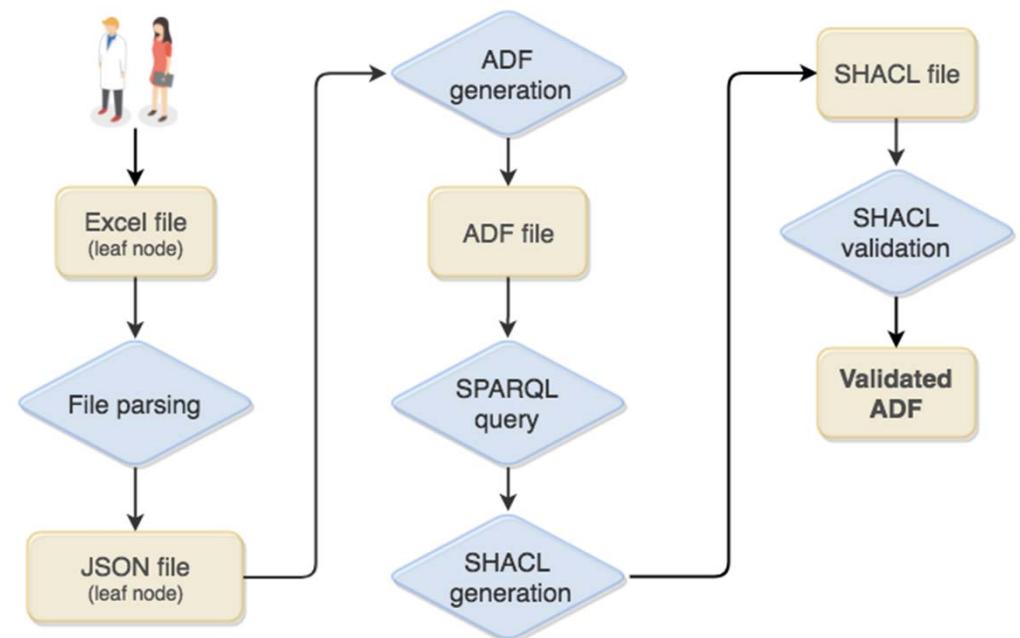


Microsoft Excel
Worksheet

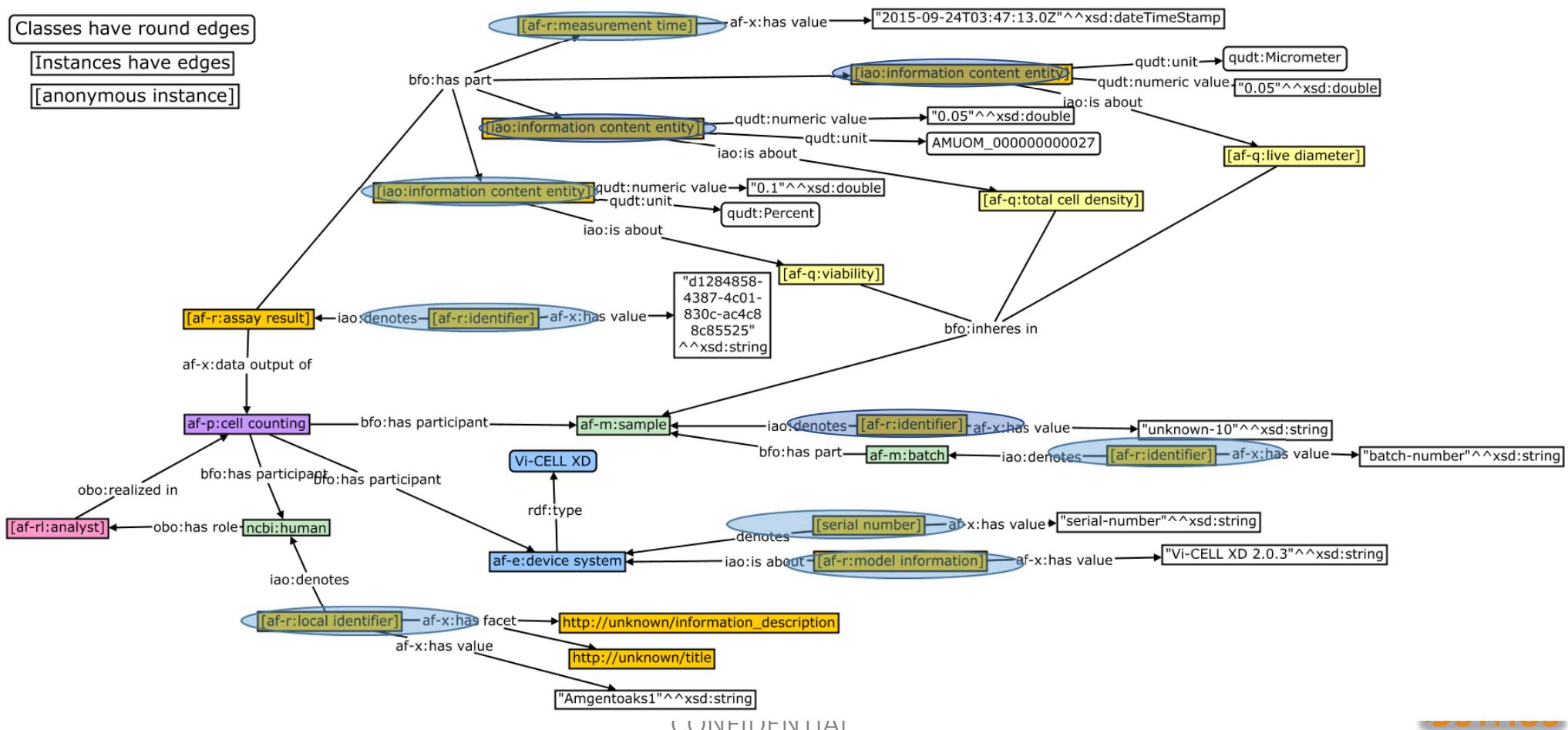
Spin: Automated ADF conversion pipeline using leaf node approach

- Starts with a human readable **Leaf Nodes Excel file**
 - Allows SME to rapidly prototype and reach agreement on data model
- End result is an ADF file that
 - is **validated by SHACL**
 - maintains **data integrity**
- Generate **relevant artifacts**
 - Data description
 - SHACL file
 - SHACL result

ADF conversion pipeline



Wes: Graph view – proposed BFO aligned model light blue anonymous nodes become explicit nodes



Observations from BFO aligned Cell Counter

Leaf node design does not preclude more complex graph structures to provide enhanced semantic meaning

- The basic model suggests that the BFO aligned structures with enhanced semantic meaning can connect to the leaf nodes at what are typically blank nodes in the ontology/graph
- Three different efforts at constructing the BFO aligned ontology for the cell counter yield approximately the same results
- Pending more testing, we have no evidence so far that the leaf node design precludes extending the data descriptor/ontology to provide richer semantic meaning

Next Steps

- Extend to RAMAN
 - Establishing model for tables in data cube
- Establish design for AKTA/HPLC/LC-UV
- Extend to new instruments based on team availability
- Extend to multiple sample, mu8ltiple injection, multiple measurement data files

Summary/conclusions

- The process works
- It maintains fidelity to the data provided by the instrument
- It is relatively efficient: potentially for a modest instrument this can be accomplished in two weeks
- There is still additional development work, but it is suggested that this work is in the main incremental

❖ **We need more teams competent in the process!!!!**

Questions/Discussion?