

Modeling Working Group: Building a path to interoperable lab data

2021 Allotrope Fall Connect

George Van Den Driessche, Biogen

Steven P. Emrick, USP

Agenda

- Why we need data models and how they support interoperability
- What is an Allotrope Data Model (ADM)
 - ADMs drive Allotrope Simple Model (ASM) generation
- Modeling Working Group Business Process



INTEROPERABILITY THROUGH DATA MODELING



Increasing volumes of data being generated in life sciences



Total amount of global healthcare data generated in 2013, compared to 2020* (in exabytes)

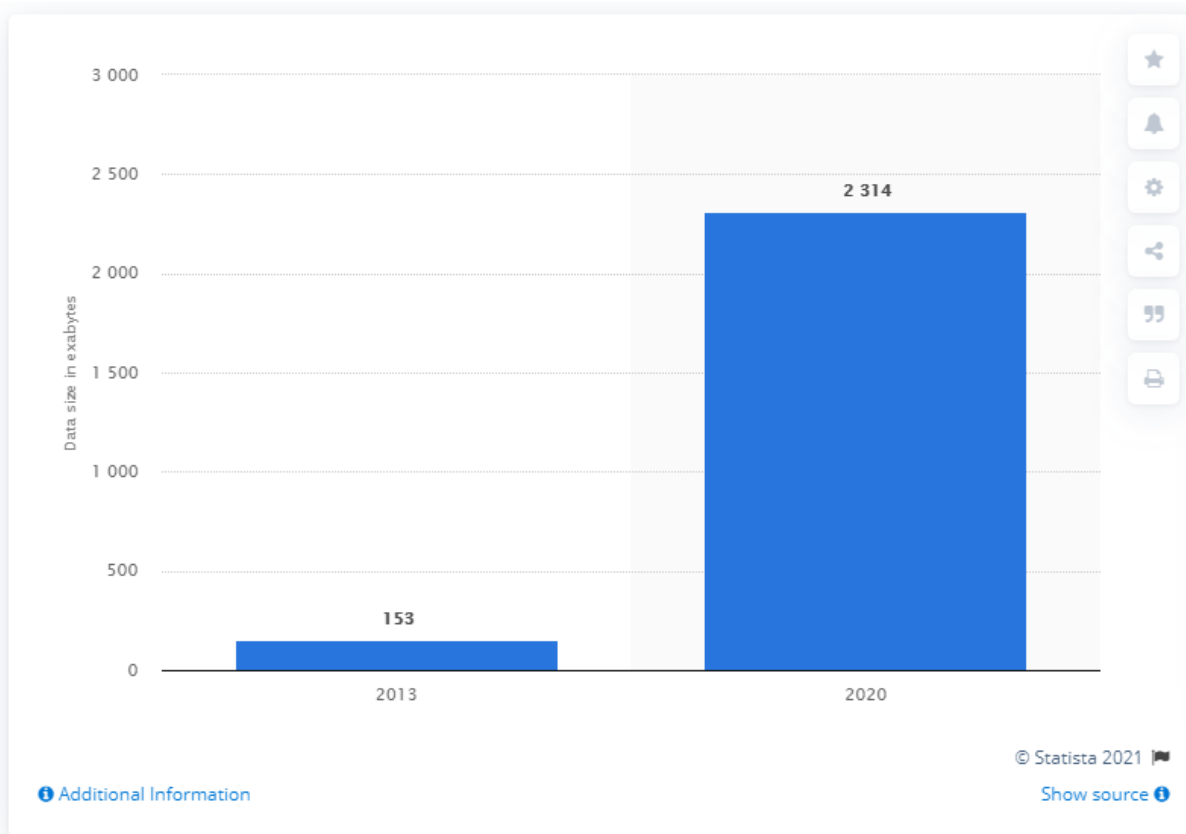


Figure 1

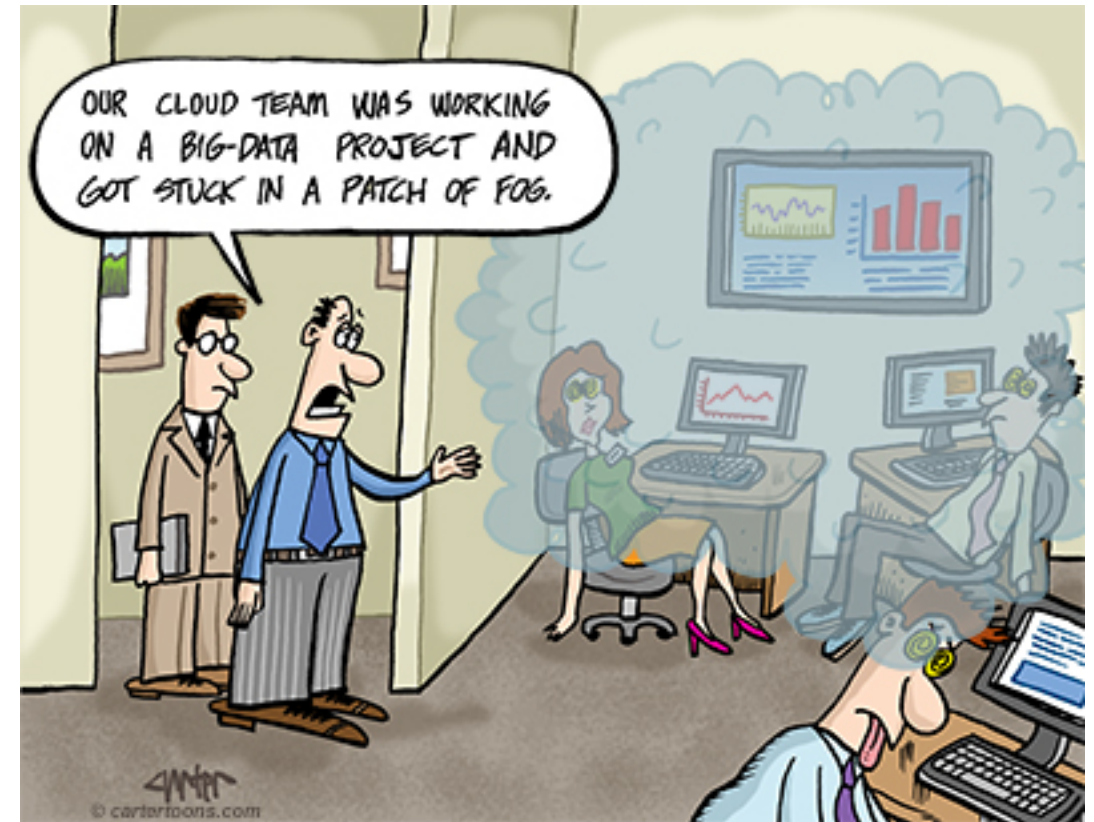



Figure 2

An example of a free text analytical method

How does an analytical chemist identify concepts that are important?

**Deconstructing an Analytical Method**


Taken From

ACS Comb Sci. 2012 Sep 10;14(9):520-6. doi: 10.1021/co300075g. Epub 2012 Aug 27. Addressing the medicinal chemistry bottleneck: a lean approach to centralized purification. Weller HN, Nirschl DS, Paulson JL, Hoffman SL, Bullock WH.


Which concepts & terms can we draw from a controlled vocabulary?

EXPERIMENTAL PROCEDURES

...A small aliquot (25 μ L) of the inbound sample was removed and diluted with 325 μ L of methanol for initial analysis. The diluted sample was injected onto an analytical LC-MS system consisting of Shimadzu LC-10 series pumps, variable wavelength UV detector (SPD-10Avp), and autosampler (SIL- 10Avp), and under control of Shimadzu ProminenceVP v 7.32.0.190 software, with a Waters model ZQ mass detector running MassLynx version 4.1 data acquisition software. Injections were made onto a Waters X-Bridge C18 column (4.6 \times 50 mm, 5 μ m particles) using mobile phase of (acetonitrile/water +10 mM ammonium acetate, with a linear gradient elution mode from 5% to 95% organic over 4 min at a flow rate of 4 mL/min. Samples were detected by UV absorbance at 220 nm and by mass spectrometry including the extracted ion chromatogram for the target (M + H)⁺ ion (456).

62

How would an algorithm identify things that are of importance to the analytical chemist?

**Deconstructing an Analytical Method**


Taken From

ACS Comb Sci. 2012 Sep 10;14(9):520-6. doi: 10.1021/co300075g. Epub 2012 Aug 27. Addressing the medicinal chemistry bottleneck: a lean approach to centralized purification. Weller HN, Nirschl DS, Paulson JL, Hoffman SL, Bullock WH.

Which concepts & terms can we draw from a controlled vocabulary?

EXPERIMENTAL PROCEDURES

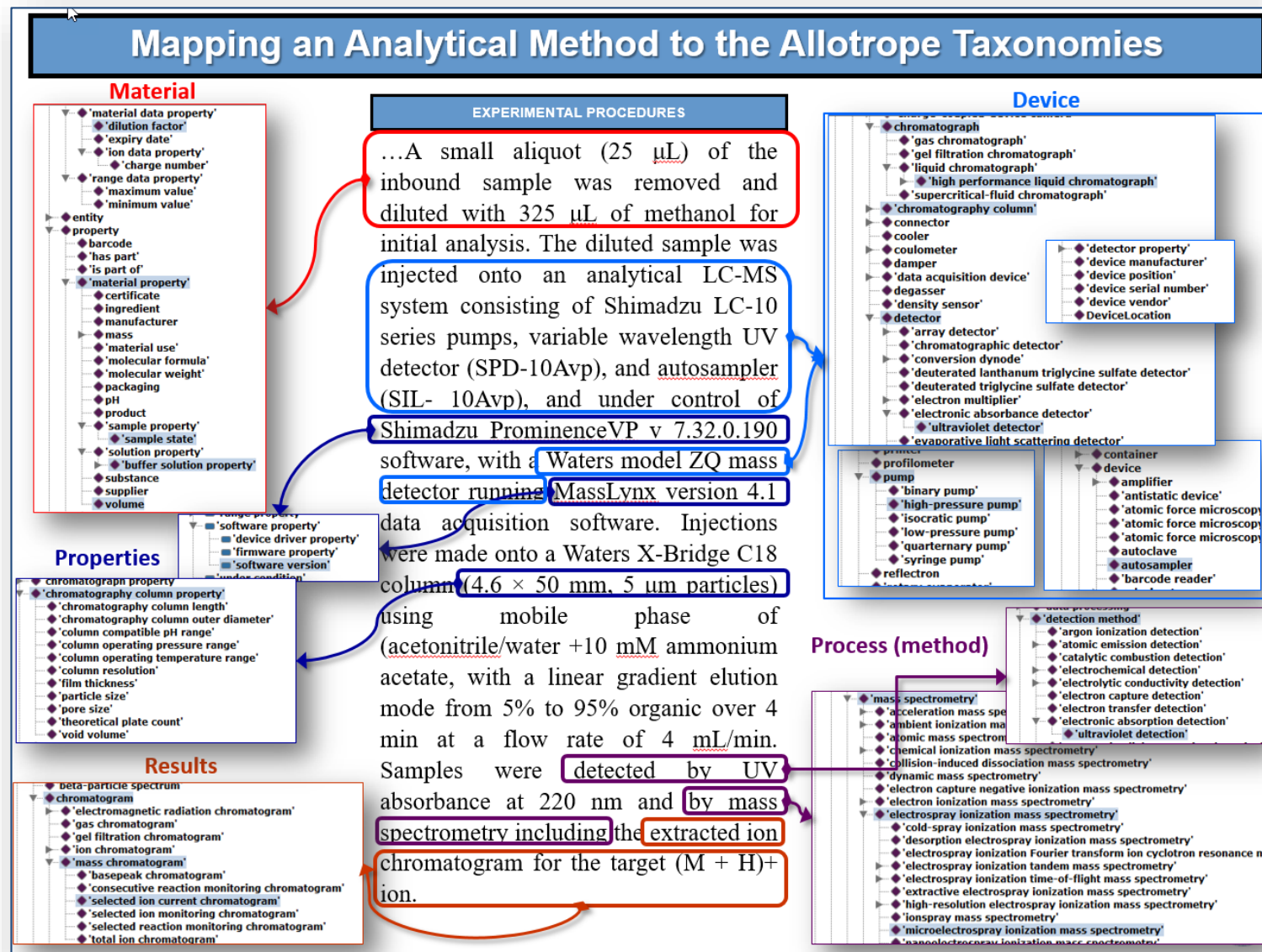
...A small aliquot (25 μ L) of the inbound sample was removed and diluted with 325 μ L of methanol for initial analysis. The diluted sample was injected onto an analytical LC-MS system consisting of Shimadzu LC-10 series pumps, variable wavelength UV detector (SPD-10Avp), and autosampler (SIL- 10Avp), and under control of Shimadzu ProminenceVP v 7.32.0.190 software, with a Waters model ZQ mass detector running MassLynx version 4.1 data acquisition software. Injections were made onto a Waters X-Bridge C18 column (4.6 \times 50 mm, 5 μ m particles) using mobile phase of (acetonitrile/water +10 mM ammonium acetate, with a linear gradient elution mode from 5% to 95% organic over 4 min at a flow rate of 4 mL/min. Samples were detected by UV absorbance at 220 nm and by mass spectrometry including the extracted ion chromatogram for the target (M + H)⁺ ion (456).

63

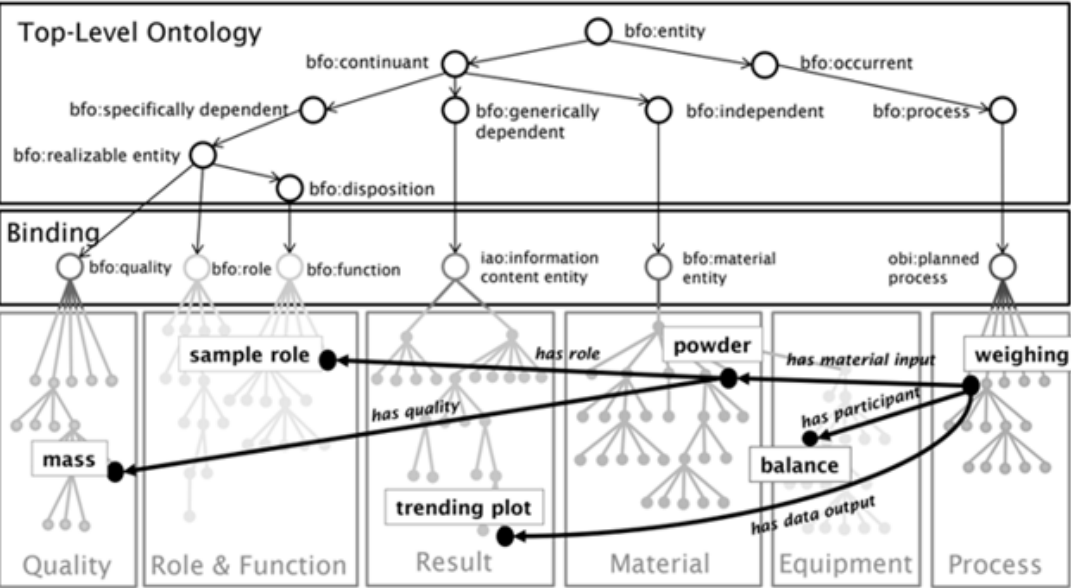
Using ontologies to organize and structure data

Classifying and Structuring data...

- ...makes us decide what things are important enough to name, classify, and relate to each other (aka data model), which...
- ...enables consistent search and retrieval of information across data sets (regardless of vendor origin), spot outliers, make decisions at scale
- Provides a common nomenclature and framework for interoperable data exchange
- Allows software and other tooling to enforce the model constraints upon data creation / entry



Classifying concepts in an Analytical Method



- AFO Process
- AFO Result
- AFO Role
- AFO Equipment
- AFO Material
- CHMO

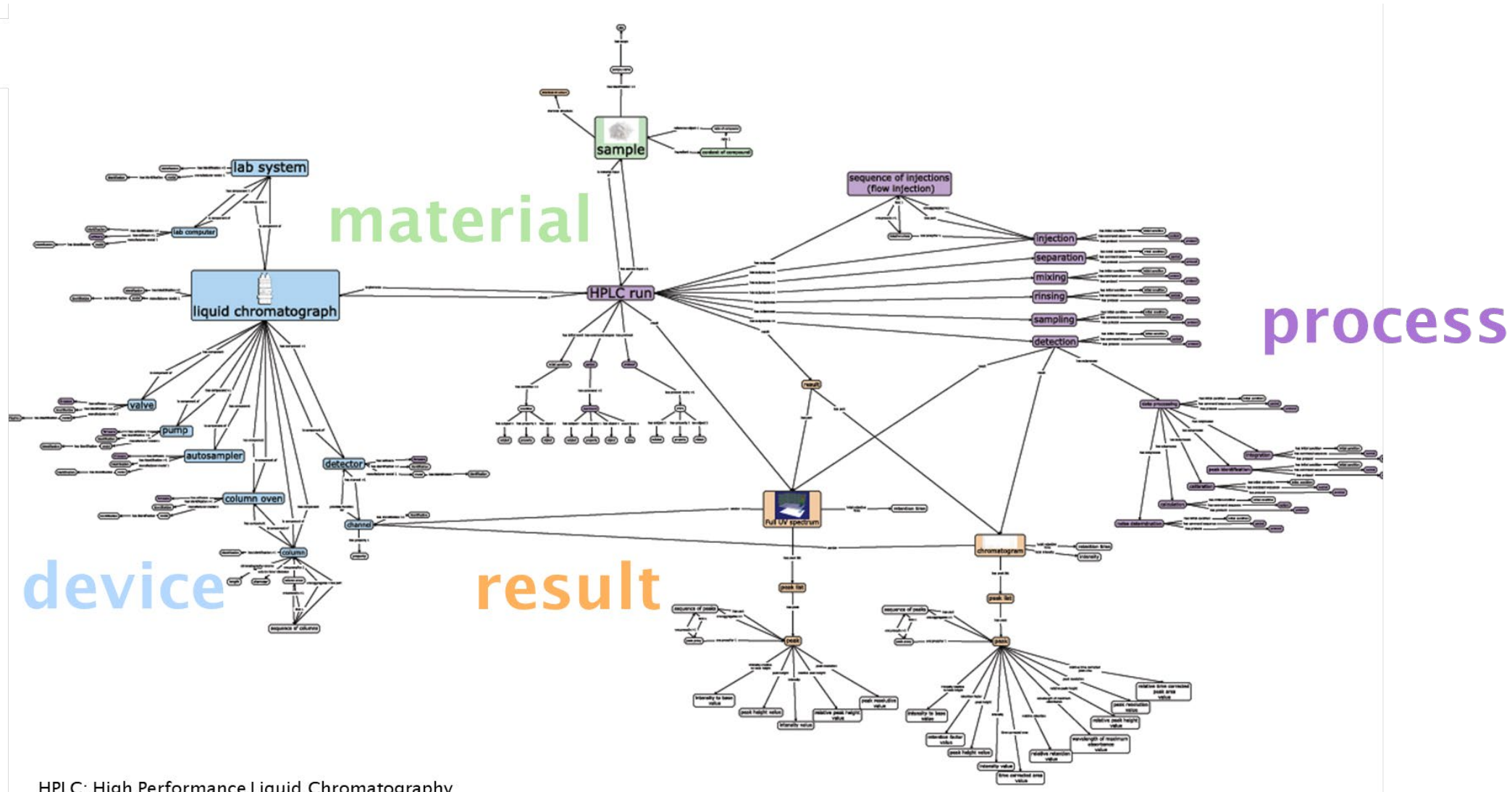
Taken From

ACS Comb Sci. 2012 Sep 10;14(9):520-6. doi: 10.1021/co300075g. Epub 2012 Aug 27. Addressing the medicinal chemistry bottleneck: a lean approach to centralized purification. Weller HN, Nirschl DS, Paulson JL, Hoffman SL, Bullock WH.

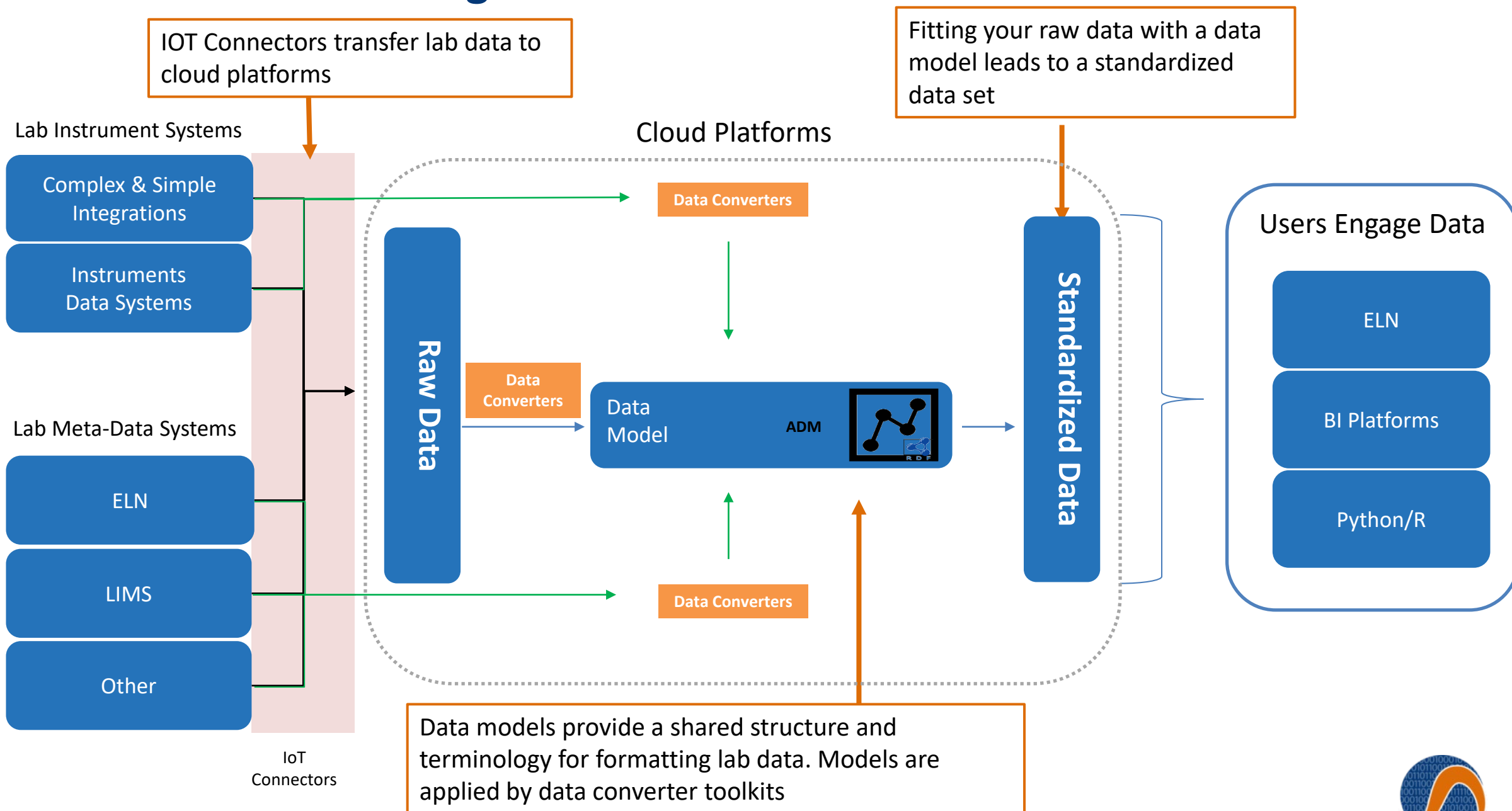
ENTITY RECOGNITION OF EXPERIMENTAL PROCEDURES

643 A small aliquot (25 μ L) of the inbound sample **SAMPLE ROLE** was removed
644 and diluted with 325 μ L of methanol for initial analysis **ANALYSIS ASSAY**. The
645 diluted sample **SAMPLE ROLE** was injected onto an analytical **LC-MS LIQUID CHROMATOGRAPHY** system **SYSTEM ROLE**
646 consisting of Shimadzu **LC-10 LIQUID CHROMATOGRAPHY** series pumps, variable **VARIABLE** wavelength **WAVELENGTH**
647 control **CONTROL SAMPLE ROLE** of Shimadzu ProminenceVP v
648 7.32.0.190 software, with a Waters **PORTION OF WATER** model **MODEL** ZQ mass detector **MASS DETECTION**
649 running **RUNNING STATE** MassLynx version 4.1 data acquisition **DATA ACQUISITION** software.
650 Sequential **SEQUENTIAL** injections **INJECTION (CHROMATOGRAPHY)** were made onto a Waters **PORTION OF WATER** X-Brid,
651 mobile phase **MOBILE PHASE** combinations (acetonitrile/water +10 mM
652 ammonium acetate, and acetonitrile/water +0.05% trifluoroacetic acid); both were run in linear gradient elution **GRADIENT ELUT**
653 5% to 95% organic over 4 min at a flow rate **RATE** of 4 mL/min.
654 Samples **SAMPLE ROLE (PREPARATION)** were detected by UV absorbance **ABSORBANCE** at 220 nm and by
655 mass spectrometry **MASS SPECTROMETRY** including the extracted ion chromatogram **ION CHROMATOGRAM**
for the target **IN-PORT ROLE** (M + H)⁺ ion (456). The two resulting

Ontology for HPLC Process, Materials, Results

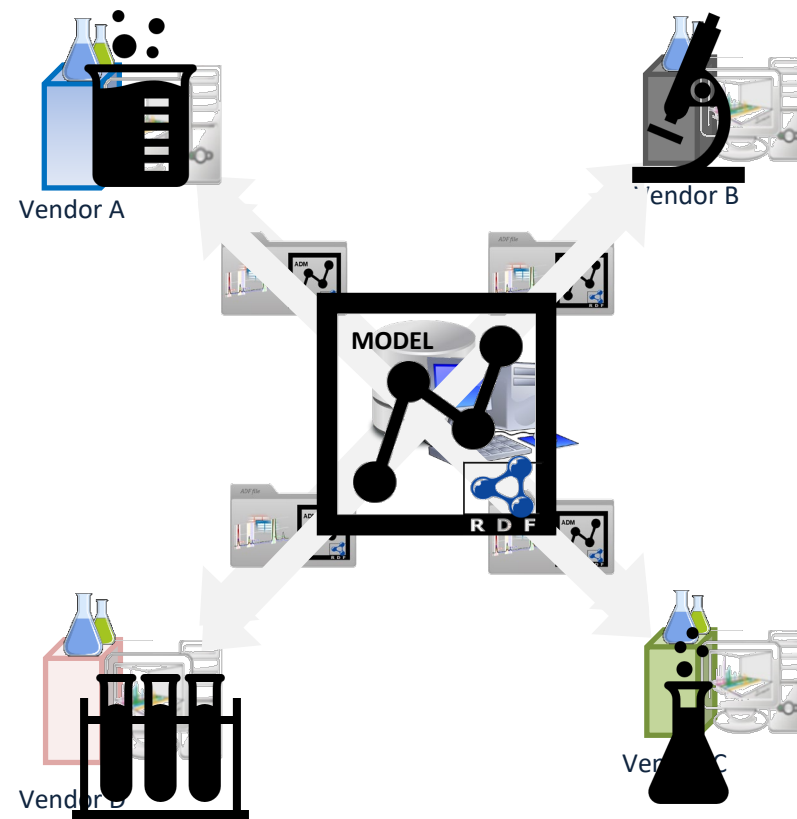
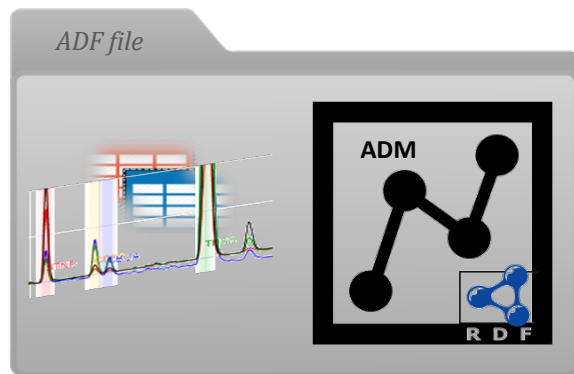


Data Models in Your Organization



What is an “Allotrope Data Model” (ADM)

- A **Model** represents one or more Use Case(s)
- The **Model** uses RDF graph to describe the Use Case(s)
- An Allotrope standardized **Model** for a laboratory analytical processes is called **ADM** (Allotrope Data **Model**)
- Heterogenous vendor systems can seamlessly exchange **ADF** files (read and write) and process its **Data** that **adheres to an associated standardized ADM**





ADM Tabular/Aggregation Benefits

Advantages

- Simple structure, reusable queries, and Excel-based creation
- Easy association of key / value pairs
- First step towards graph description
- Leafnode classes have better semantics than textually annotating
- Classes can be created by concatenating terms
- Collapsing contextual information into single node
- Structured aggregations allow grouping of related metadata

Drawbacks

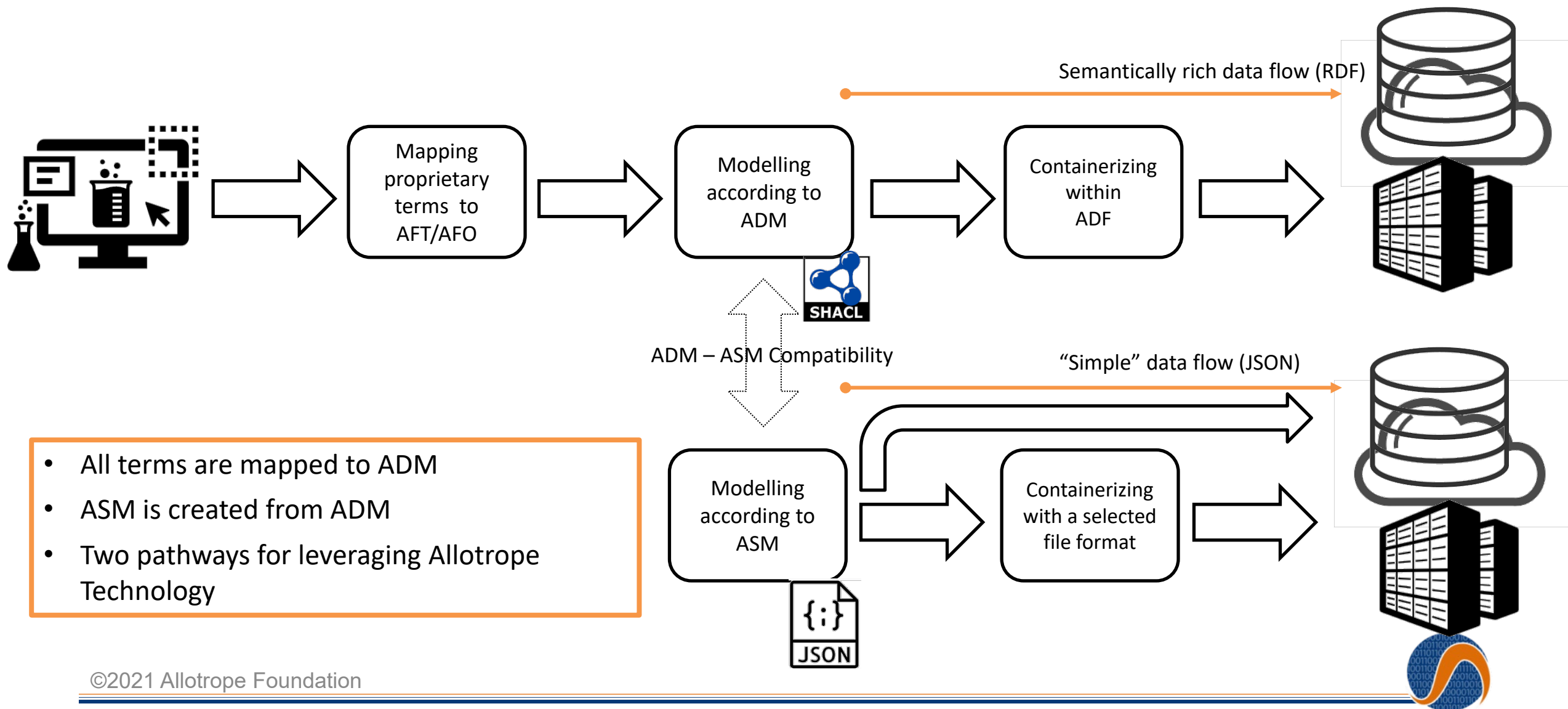
- Provide limited context until connected back to AFO
- Disconnected nodes are not easily merged together
- Aggregations are easier to merge but still not trivial
- Less machine-understandable due to limited semantics & context
- Difficult to extend if details of leafnodes are to be described
- Data cannot be interrogated by a single query easily
- Multiple queries needed to understand leafnode values and hierarchy

Parameter Type	Parameter	Parameter prefLabel	Parameter Allotrope URI	Required	Parameter Units of measure URI	Example Parameter Value	Parameter Unit Symbol	Parameter Unit Qudt URI
metadata	Measurement ID	measurement identifier	http://purl.allotrope.org/ontologies/result#AFR_0001121	N	xsd:string	413befdd-c7e2-4edd-9e9b-06cf1cb0283f		
metadata	RunDate	measurement time	http://purl.allotrope.org/ontologies/result#AFR_0000952	Y	xsd:dateTime	2015-09-24T03:47:13.001Z		International System of Units
metadata	Operator	analyst	http://purl.allotrope.org/ontologies/result#AFR_0001116	N	xsd:string	operator-10		
metadata	Sample ID	sample identifier	http://purl.allotrope.org/ontologies/result#AFR_0001118	Y	xsd:string	sample-123		
metadata	Machine ID	equipment serial number	http://purl.allotrope.org/ontologies/result#AFR_0001119	N	xsd:string	serial-number-XYZ		
metadata	Batch ID	batch identifier	http://purl.allotrope.org/ontologies/result#AFR_0001120	N	xsd:string	batch-number-100		
Results Data	conductivity	conductivity	http://purl.allotrope.org/ontologies/result#AFR_0001587	Y	xsd:double	273000	S/m	http://purl.allotrope.org/ontology/qudt-ext/unit#SiemensPerMeter
Results Data	Temperature	temperature	http://purl.allotrope.org/ontologies/result#AFR_0001584	N	xsd:double	28.6	degC	http://qudt.org/vocab/unit#DegreeCelsius

Example: Conductivity Measurement Model (.xlsx file)

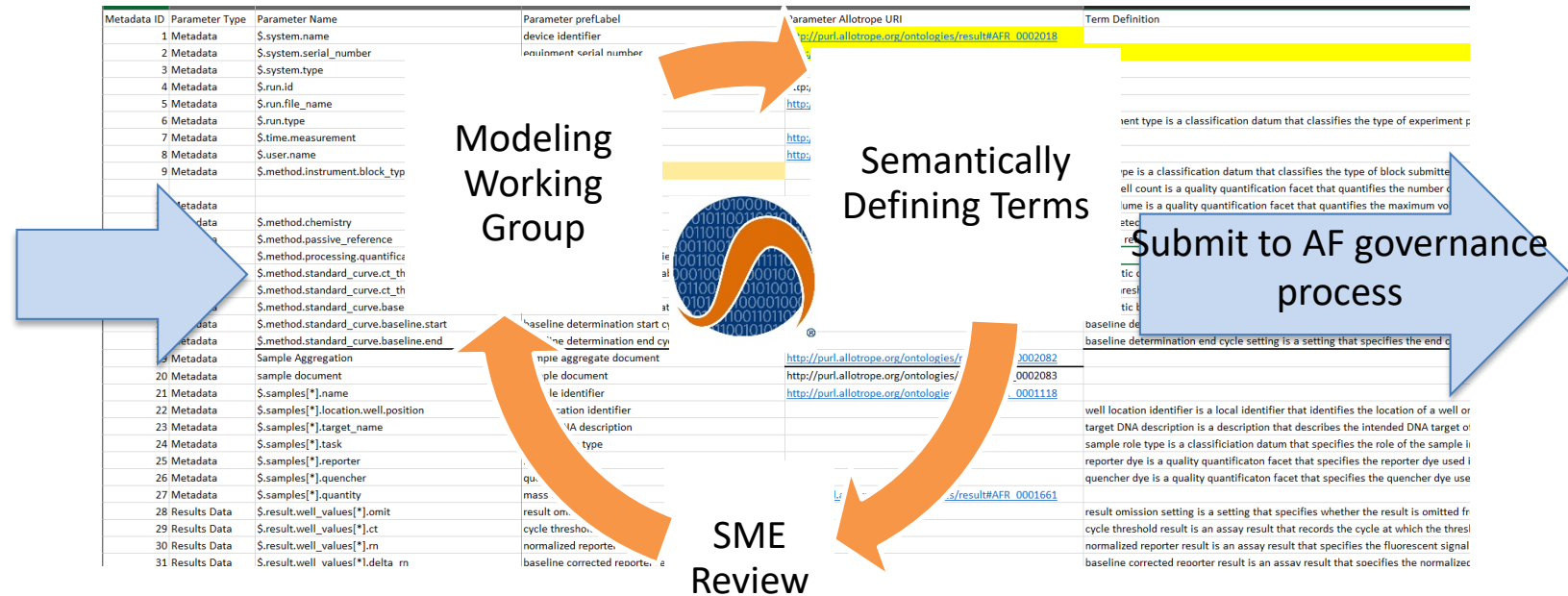


Data Flow Using the Simple Model vs ADM



WG Business Process for Drafting a Tabular Model

Parameter Type	Path in IDS	Term Definition (defined by SME)
metadata	\$.system.name	
	\$.system.serial_number	
	\$.system.type	
	\$.run.id	
	\$.run.file_name	Instrument Output file name
	\$.run.type	Quantitative or Qualitative experiment
	\$.time.measurement	
	\$.user.name	Analyst
	\$.method.instrument.block_type	Size and Volume of PCR plate
	\$.method.chemistry	Chemistry for PCR Detection
	\$.method.passive_reference	Reference dye used for normalization
	\$.method.processing.quantification_cycle_method	Quantification Measure
	\$.method.standard_curve.ct_threshold.automated	Automatic CT Threshold Determination
	\$.method.standard_curve.ct_threshold	CT Threshold
	\$.method.standard_curve.baseline.automated	Automatic baseline determination
	\$.method.standard_curve.baseline.start	Cycle where baseline determination starts
	\$.method.standard_curve.baseline.end	Cycle where baseline determination ends
	\$.samples[*].name	Sample name
	\$.samples[*].location.well.position	Plate well Location
	\$.samples[*].target_name	Probe target for amplification
	\$.samples[*].task	Purpose of sample
	\$.samples[*].reporter	Dye used for Reporter
	\$.samples[*].quencher	Dye used for Quenching
	\$.samples[*].quantity	Preetermined std quantity or backcalculated concentration
	\$.result.well_values[*].location.well.position	Plate well Location
	\$.result.well_values[*].omit	was result omitted from internal calculation
	\$.result.well_values[*].ct	Cycle threshold value, which cycle was the threshold passed
	\$.result.well_values[*].rn	Normalized reporter value, sample signal/passive reference signal
	\$.result.well_values[*].delta_rn	Sample Rn - Baseline Rn



Prioritized Instrument Term List

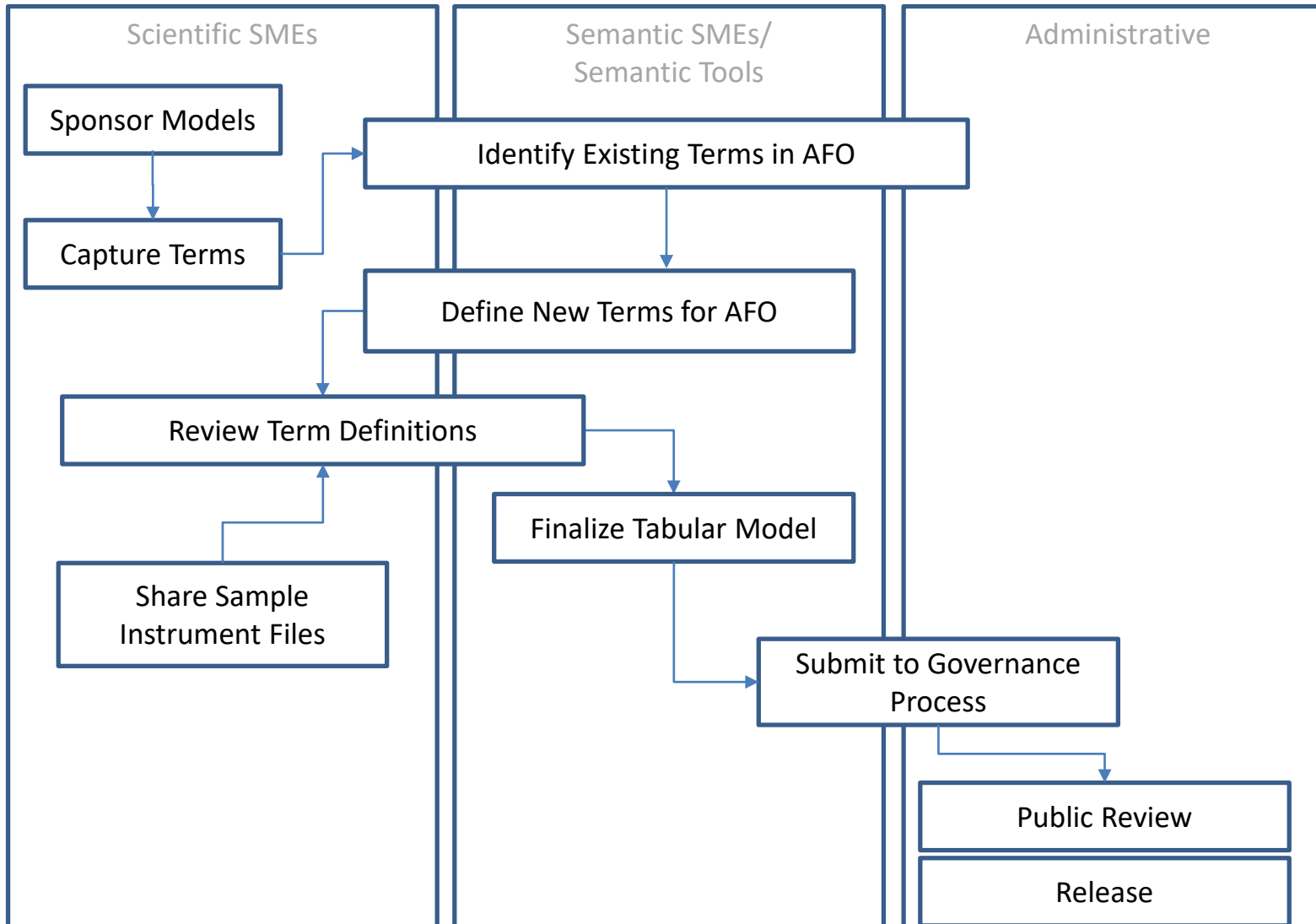
- Company rep sponsors instrument for model development
- SMEs draft a priority term list
- Define model use case

Modeling WG Business Process

- Prioritized term list submitted to WG
- WG members & semantic engineers formally review term definitions & assign to AFO
- Draft model shared with partner SMEs for formal review



Overview of Modeling WG Roles



- Collaborative partnership drives success
- Scientific SME's
 - Provide domain knowledge
 - Gain semantic knowledge
- Semantic SMEs
 - Provide semantic expertise
 - Gain domain knowledge



Data Model and Ontology Governance Process

The process for developing and integrating new Allotrope Data Models (ADM) and new ontology terms into the Allotrope Foundation Ontologies (AFO) consists of five phases, from creation of the initial group to develop the new ADM/AFO through its official inclusion as an Allotrope Recommendation.

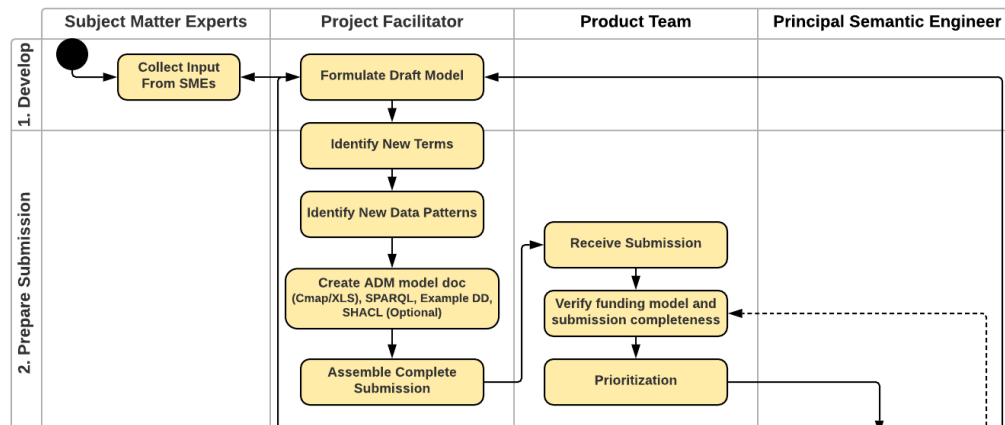
- Data Model and Ontology Development
 - **Phase I:** Model Ideation and Development
 - **Phase II:** Formalization of Draft Model and/or Additional Taxonomy Terms
- Draft Model Governance
 - **Phase III:** Draft Model Review
- Public Review
 - **Phase IV:** Public Review Procedure
- Incorporation of Model as a Formal Allotrope Recommendation
 - **Phase V:** Periodic Release of ADM and AFO



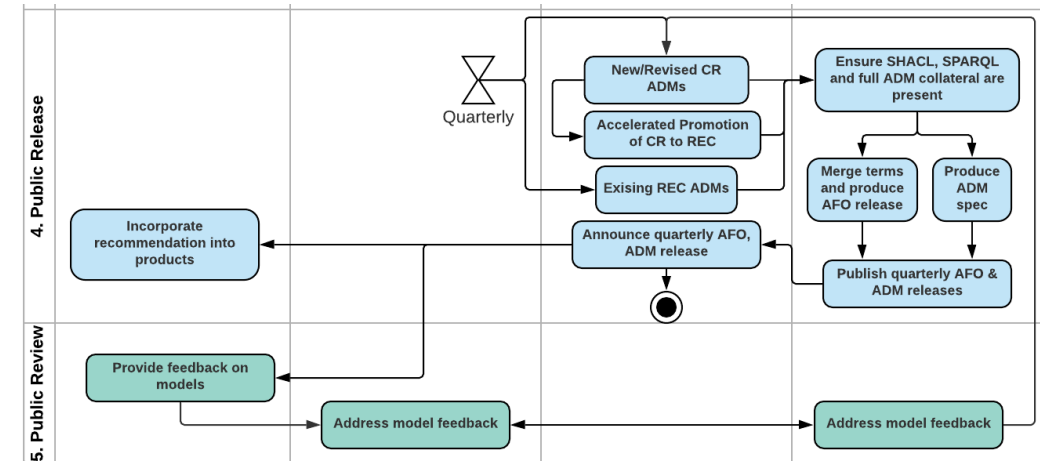
Data Model and Ontology Governance Process

- **Phase I: Model Ideation and Development**
- **Phase II: Formalization of Draft Model and/or Additional Taxonomy Terms**

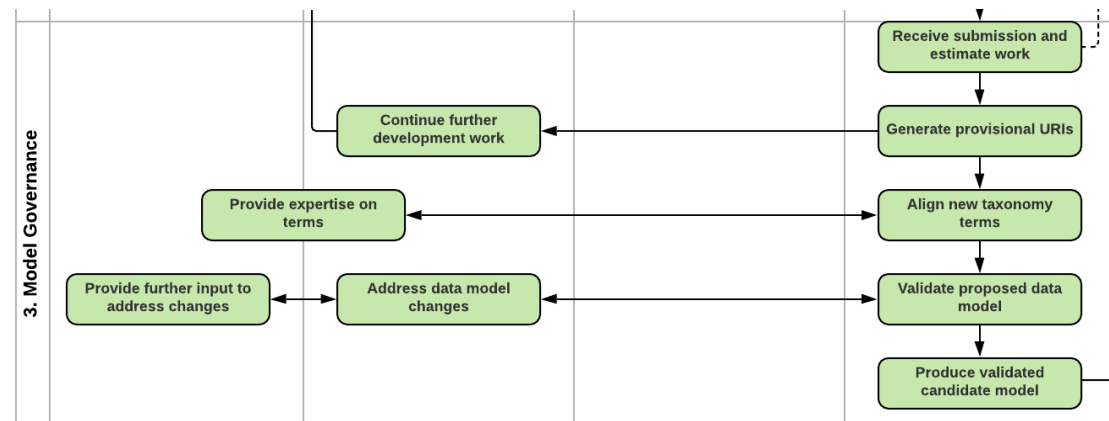
ADM GOVERNANCE PROCESS (VERSION 2)



- **Phase IV: Public Review Procedure**
- **Phase V: Periodic Release of ADM and AFO**

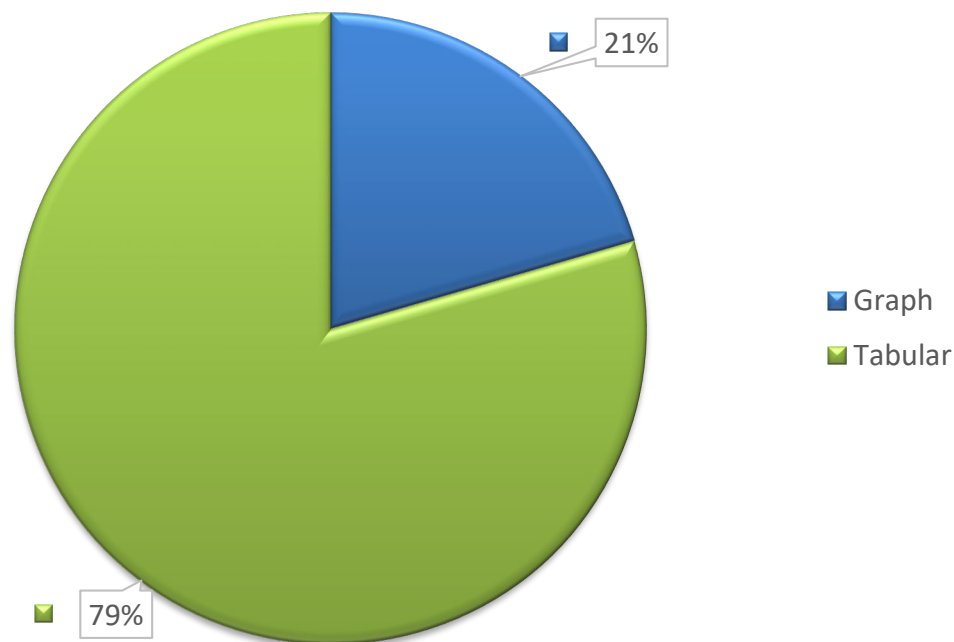


Phase III: Draft Model Review



Over 80% of ADMs are tabular, and do not require a full graph model for standardized data representation

Breakdown of Graph vs Tabular ADM Models



<https://www.allotrope.org/product-releases>

Model	Type	Maturity	Availability (*targeted)	Status
Automated Reactors - PAT (Process Analytical Technology)	Tabular Graph	REC	2021/03 (*2021/09)	Released & In Review In Draft
Balance	Tabular	REC	2020/06	Released
Blood Gas Analyzer	Tabular	REC	2020/03	Released
Bulk Density	Tabular	REC	2020/12	Released
Calibration	Graph	REC	2021/03	Released
Cell Counting	Tabular	REC	2020/06	Released
FTIR	Tabular	REC	2021/03	Released



How to Join

- Modeling WG meets weekly on Wednesdays at 11 AM ET (4 PM BST, 5 PM EET)
- Contact Ben Woolford-Lim to be added to the mailing list serv
 - benjamin.woolford-lim@allotrope.org
- Product Releases - <https://www.allotrope.org/product-releases>
- ADM Governance Process - https://community.allotrope.org/resources/reference/semantic/governance/afo_adm_governance_process/
- The official repository for the ADM artifacts: <https://gitlab.com/allotrope/adm>

Questions??



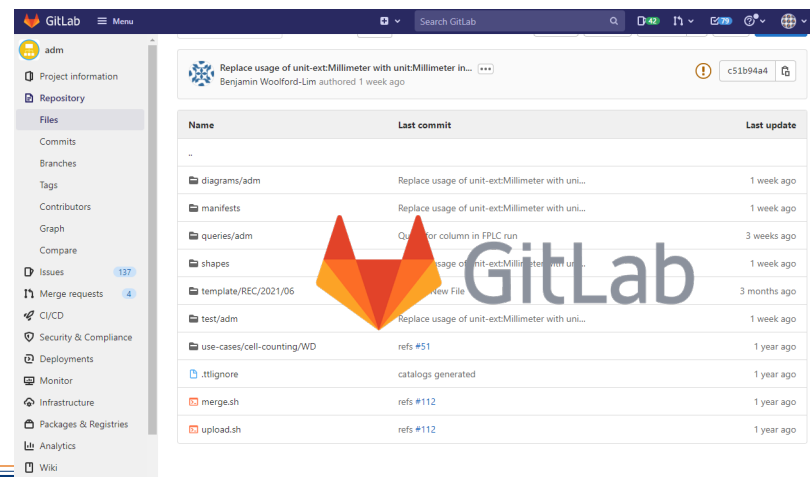






Supporting Info








Modeling WG:

Where are the working artifacts?

- The official repository for the ADM artifacts: <https://gitlab.com/allotrope/adm>
- Repository structure:
 - Tabular model Excel spreadsheets: <https://gitlab.com/allotrope/adm/-/tree/develop/purl/diagrams/adm>
 - SHACL Shape files: <https://gitlab.com/allotrope/adm/-/tree/develop/purl/shapes>
 - Manifest files: <https://gitlab.com/allotrope/adm/-/tree/develop/purl/manifests>
 - Test, data Instance files (ADF and RDF): <https://gitlab.com/allotrope/adm/-/tree/develop/purl/test/adm>
 - SPARQL queries files (.rq): <https://gitlab.com/allotrope/adm/-/tree/develop/purl/queries/adm>



 Replace usage of unit-ext:Millimeter with unit:Millimeter in...   c51b94a4 
Benjamin Woolford-Lim authored 1 week ago

Name	Last commit	Last update
..		
diagrams/adm	 Replace usage of unit-ext:Millimeter with uni...	1 week ago
manifests	 Replace usage of unit-ext:Millimeter with uni...	1 week ago
queries/adm	 Query for col... C run	3 weeks ago
shapes	 Replace usage of unit-ext:Millimeter with uni...	1 week ago
template/REC/2021/06	Upload New  Shape	3 months ago
test/adm	 Replace usage of unit-ext:Millimeter with uni...	1 week ago
use-cases/cell-counting/WD	refs #51  TURTLE	1 year ago
.ttlignore	catalogs gene	1 year ago
merge.sh	refs #112	1 year ago
upload.sh	refs #112	1 year ago

GitLab ADM repository

- Tabular model Excel spreadsheets
- Manifest files
- SPARQL queries files (.rq)
- SHACL Shape files
- Test, data Instance files (ADF and RDF)