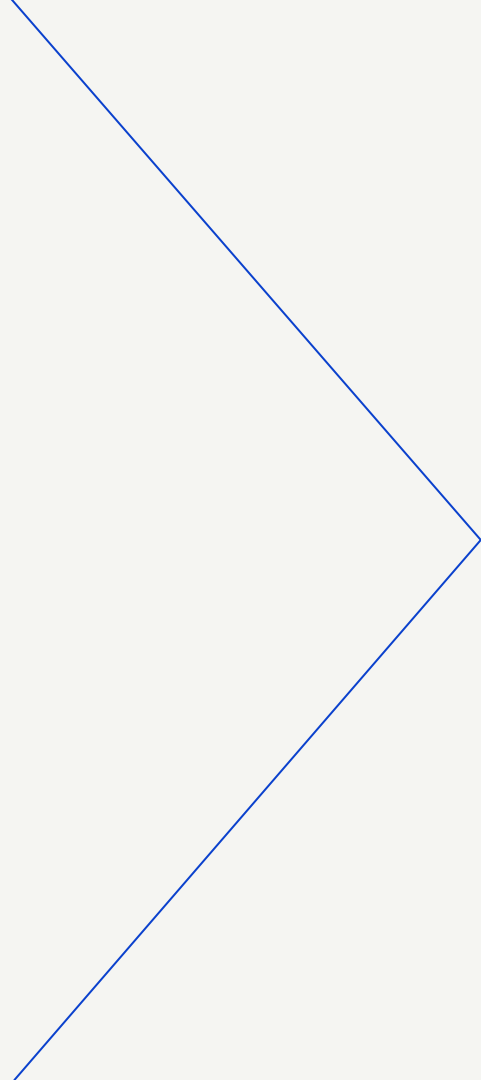


Breaking down the data standards silos

A new metadata ontology for data analysis derivations

Nelia Lasierri and Chris Price, Roche

Agenda

- 
- A large blue arrow pointing from the left towards the agenda items, formed by two diagonal lines meeting at a point on the left edge of the slide. A vertical blue line runs parallel to the right edge of the arrow's shaft.
- 1. The value of Data, and the impact of FAIR
 - 2. How did we start Orchid Garden?
 - 3. Ontology Development and Assessment
 - 4. Results
 - 5. Next Steps and Outlook

The value of Data, and the impact of FAIR

Towards workflow optimization and automation

Transforming data into patient benefit

Maximizing the value that we obtain from our data: faster and better insights

Our Vision, our goals

- Develop more and better medicines
- Drive personalized healthcare
- Reduce time to find new treatments
- Faster and more effective insights

Speed to Insight requires
optimize delivery & help enhance
understanding of a molecule or
therapy

Key enablers for success

- Data ready for insights discovery
- Improve management of data assets
- Connected and harmonized data at the center



Digital applications can help us with Data Science activities enabling automation of business workflows, knowledge discovery and sharing data at scale

FAIR data

FAIR - a catalyst for digital transformation

- “FAIR data shared” concept was introduced (4-5) years ago
 - Unlocking the possibilities of combining and sharing data across the organization
- FAIR is a recognized strategy to enable better data management activities and increase quality
 - FAIRification of processes, systems, data and FAIR culture
- Adopting FAIR principles enable to describe the data in a way digital solutions “understand” the data and can help us to share, connect, improve data quality and maximize efficiency
- FAIR implementations need to be aligned and interoperable

FAIR By Design

Perspectives

FAIR implementation

- **Identifiers**
 - Unique, persistent and decoupled of business knowledge and systems
- **Data Standards**
 - Ensure common understanding
 - Ontologies
- **Rich metadata** how, what, why, by whom
 - Gives the **CONTEXT**
- **Access and re-use**
 - Standard protocol

“Making the FAIR way, the easy way”

- **Business oriented questions**
 - Target business workflows
- **User Centric Design**
 - Easy to use
- **Rich metadata**
 - Collected once
- **Prospective FAIRification**
 - Avoid retrospective tagging

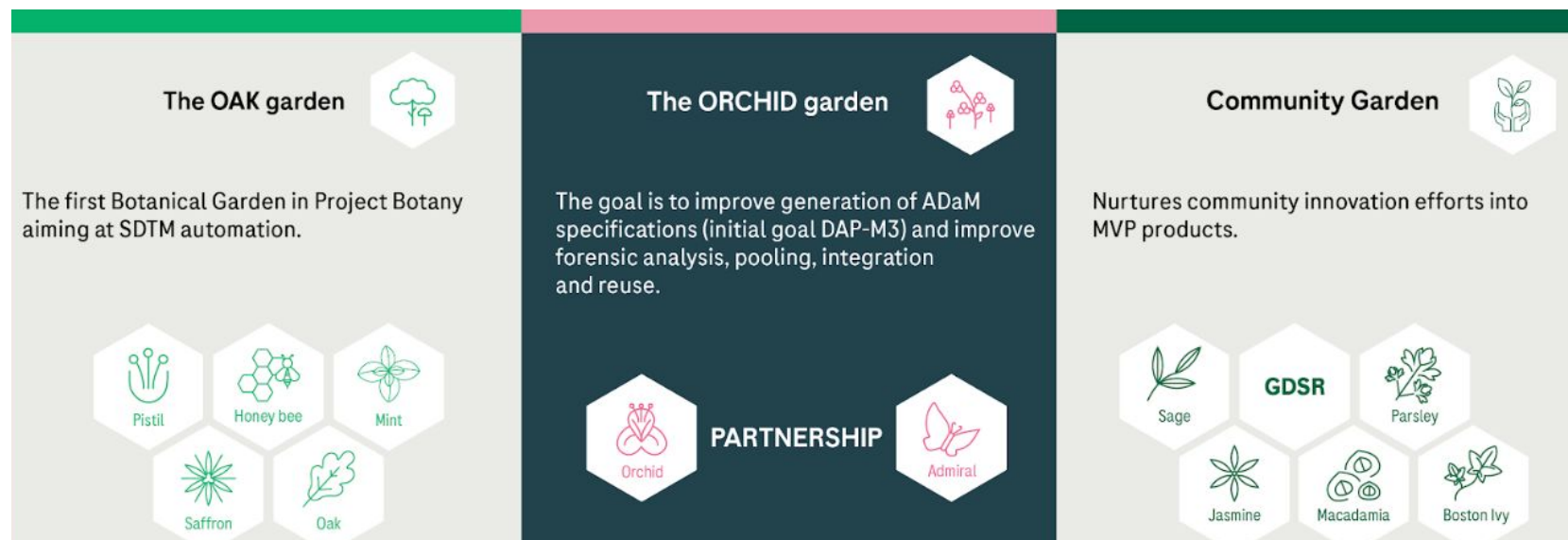
Project Botany

FAIRification of workflows along the drug development lifecycle

User centric - Optimizes and FAIRify study conduct workflows through reuse and automation workflows

- Facilitate standards adoption, Data harmonisation and FAIR best practices

Aligned FAIRification Strategy - Identifier strategy, ontologies, tools and management of digital context



Orchid Garden

Enables FAIRification of analysis dataset specification

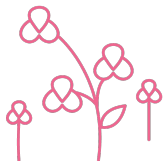
Analysis Specs

The **analysis dataset specification** is a fundamental document for a clinical data scientist preparing to analyse data

- Serve as documentation and programming instruction for analysis datasets
- Required as part of the submission to Health Authorities (**define.xml**)
- Instrumental in making data FAIR
- Enable re-use and understanding of how the datasets were created

The goal of Orchid Garden is to **optimize design, delivery and understanding of ADaM dataset** generation addressing key challenges involved in the creation and usage of the analysis specifications.

The Orchid Garden solutions address these challenges and optimises management of specifications by capturing **FAIR machine readable metadata**



How did we started Orchid Garden?

Use Cases and initial steps

Challenges with the analysis specifications

Generation, Integration, re-use and traceability

Generation

- **Complexity**
 - Manual process, error prone
- **Code relationship**
 - Guideline not used directly in the code
- **Multiple Use cases**
 - Technical and Functional
 - Human and machine readable

Integration, re-use and traceability

- **Harmonisation**
 - Complex alignment across Molecule/Project
- **Findability**
 - No queryable content machine readable
 - Lack of central repository
- **Traceability**
 - Understanding provenance is a manual process

Generation of analysis documentation

Enabling automation, inheritance and improving quality



Data Scientist

Scenario:

As part of generating new ADaM datasets, I need to generate the analysis dataset specification.

Goal / workflow to be optimized:

- All content available as machine readable to provide the opportunity to automatically generate human and machine readable renderings of specifications (e.g. define.xml).
- Improve the data and metadata alignment to support health authority submission related processes.
- Improved relationships between specification contents to reduce manual input
- Improved support of molecule consistency by the enablement of inheritance from molecule-level specifications as opposed to copying of content.
- Increased efficiency of data pooling activities by efficient identification of differences between individual studies.

Questions:

- What is the algorithm/choice of algorithm for an individual derivation?
- What are the ADaM variables and where necessary the ADaM Parameters that are created for this derivation?
- What derivations are prerequisites in order to perform another derivation?
- What source data (SDTM) is required for this derivation?
- What functions do I need to generate this dataset?

Target Curation and metadata traceability

Understanding metadata and data transformation



Data Scientist

Scenario:

To create a new data pool, I need to be able to understand how analysis derivations were calculated in multiple studies and any gaps in pooling these studies. Based on the analyses of interest I need to be able to target SDTM domains, tests and variables for curation.

Goal / workflow to be optimized:

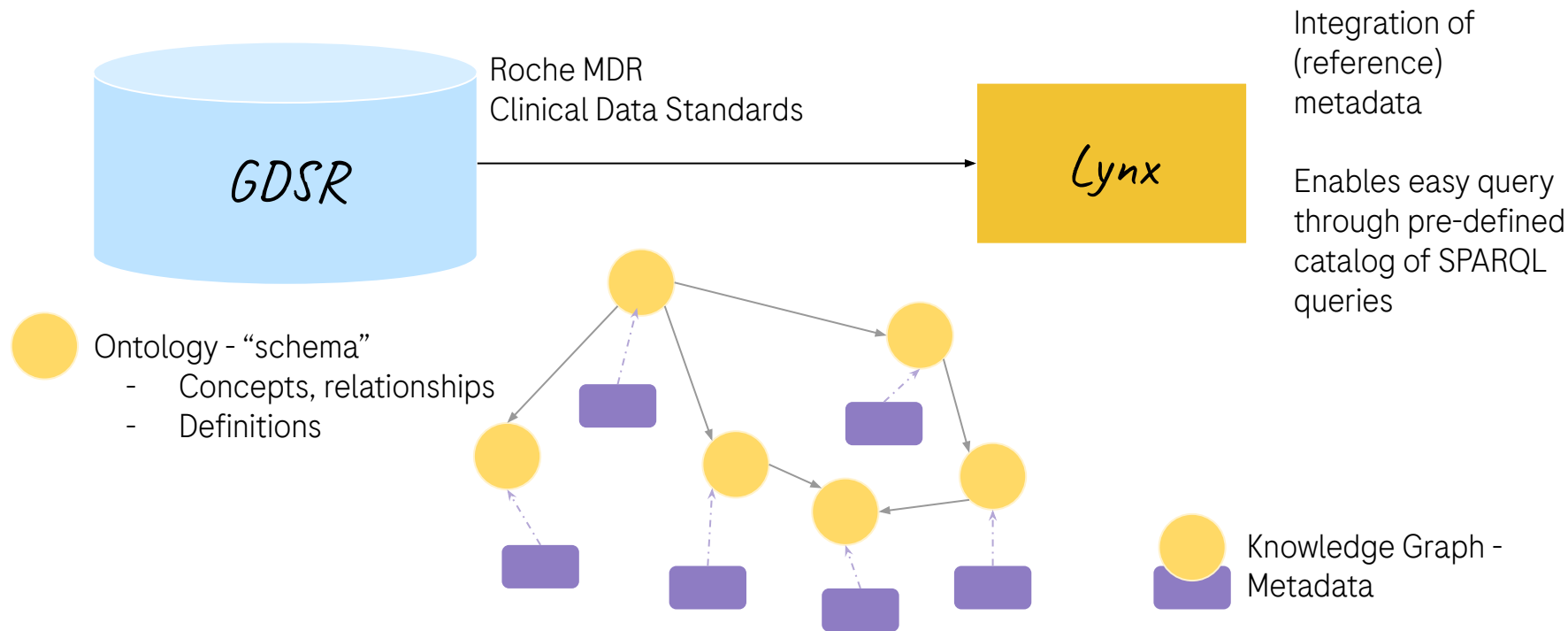
- Enable querying and identification of implemented content across studies.
- Improved consistency of specifications across both a molecule and therapeutic area.
- Introduce the automated ability to identify specific datasets and variables that need to be curated in order to perform a derivation to target curation efforts where they bring the most value.
- Improved capability to conduct beginning-to-end traceability for analyses.

Questions:

- What is the data (SDTM variables including identifiers) that impact the calculation of a derivation?
- What is the algorithm/choice of algorithm for an individual derivation?
- What derivations (and consequently endpoints) were created for a clinical study?
- What derivations could be created (based on the available source data) for a clinical study?

Introductions

Key elements for our work



Initial Steps

“The Experiment 0”

Team Set Up

Form an interdisciplinary team with expertise in Semantic Technologies, SPARQL, CDISC and the analysis dataset specification process

Ontology and Knowledge Graph Development

Design an ontology to explicitly connect SDTM variables and ADaM variables via derivations

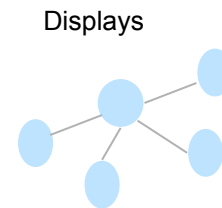
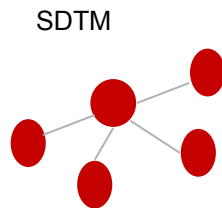
- Overall Survival
- Diastolic Blood Pressure Change from Baseline.

Publication in Lynx

Publish the created ontology and knowledge graph in the GDSR and enable access to the Data Scientist through the Lynx (Lab)

Ontology Development and Assessment

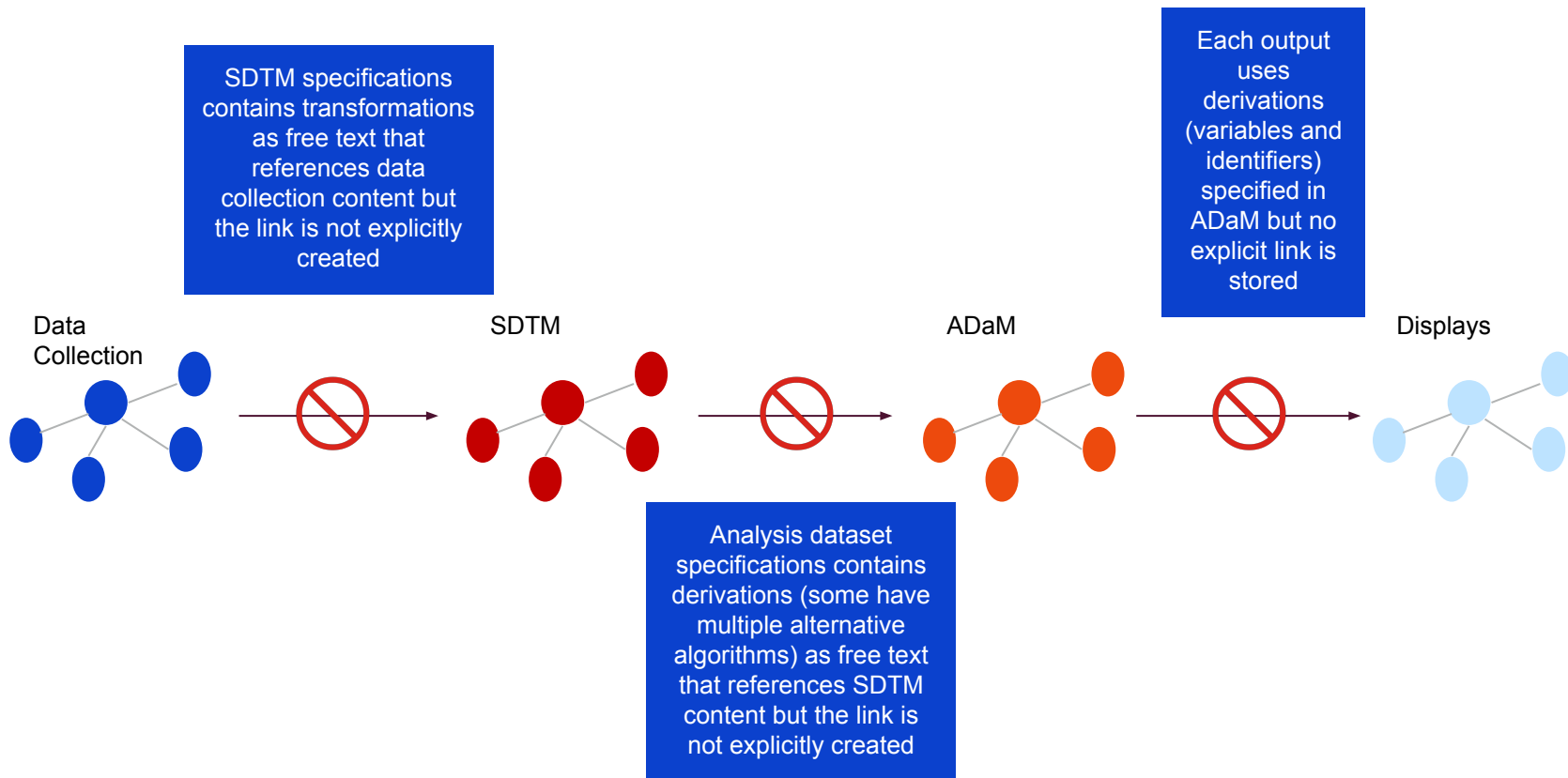
Current Data Standards Models



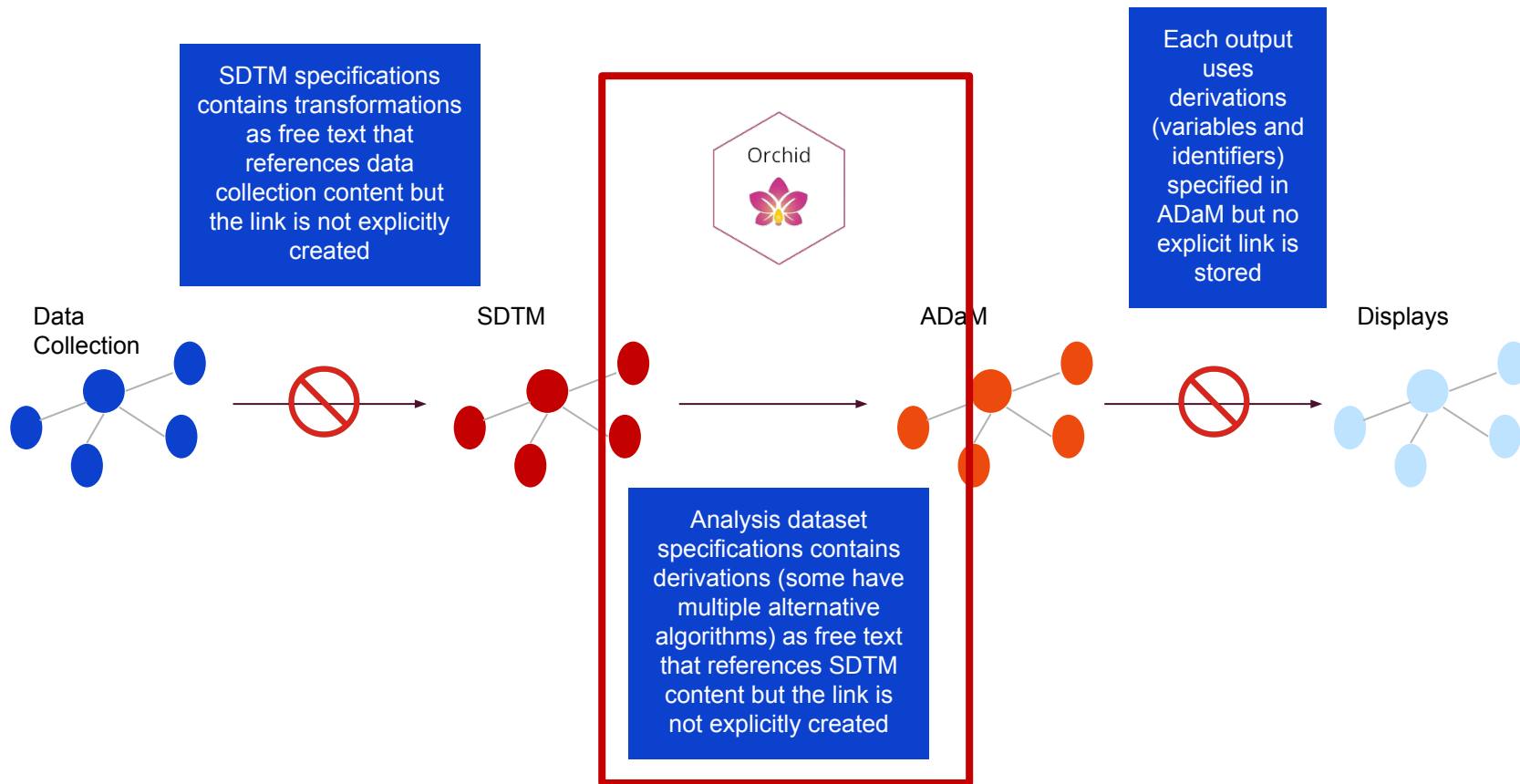
Current Data Standards Models



Current Data Standards Models



Current Data Standards Models



Ontology Development

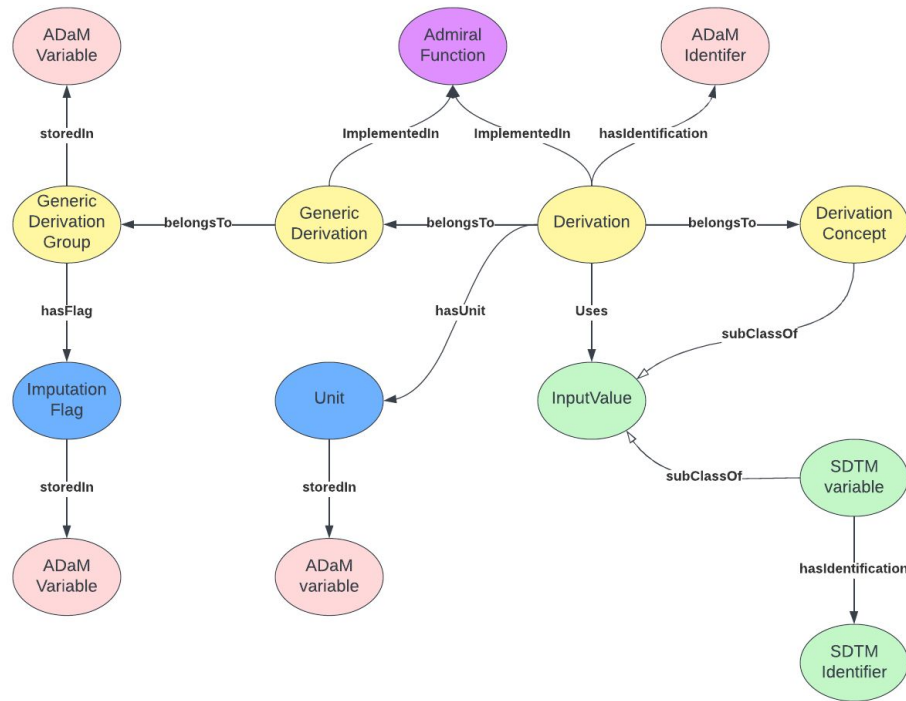
Key Competency Questions

- What are all the SDTM variables that impact a ADaM derivation?
- What are the Admiral functions needed for generating derivation X?
- Identify all ADaM variables and parameters required for a derivation



Requirements

- Defining the input for a derivation
- Defining the output of a derivation
- Defining a link to the implementation of the derivation
- Support of multiple algorithms to derive the same analysis concept
- Provision of persistent URIs for derivations

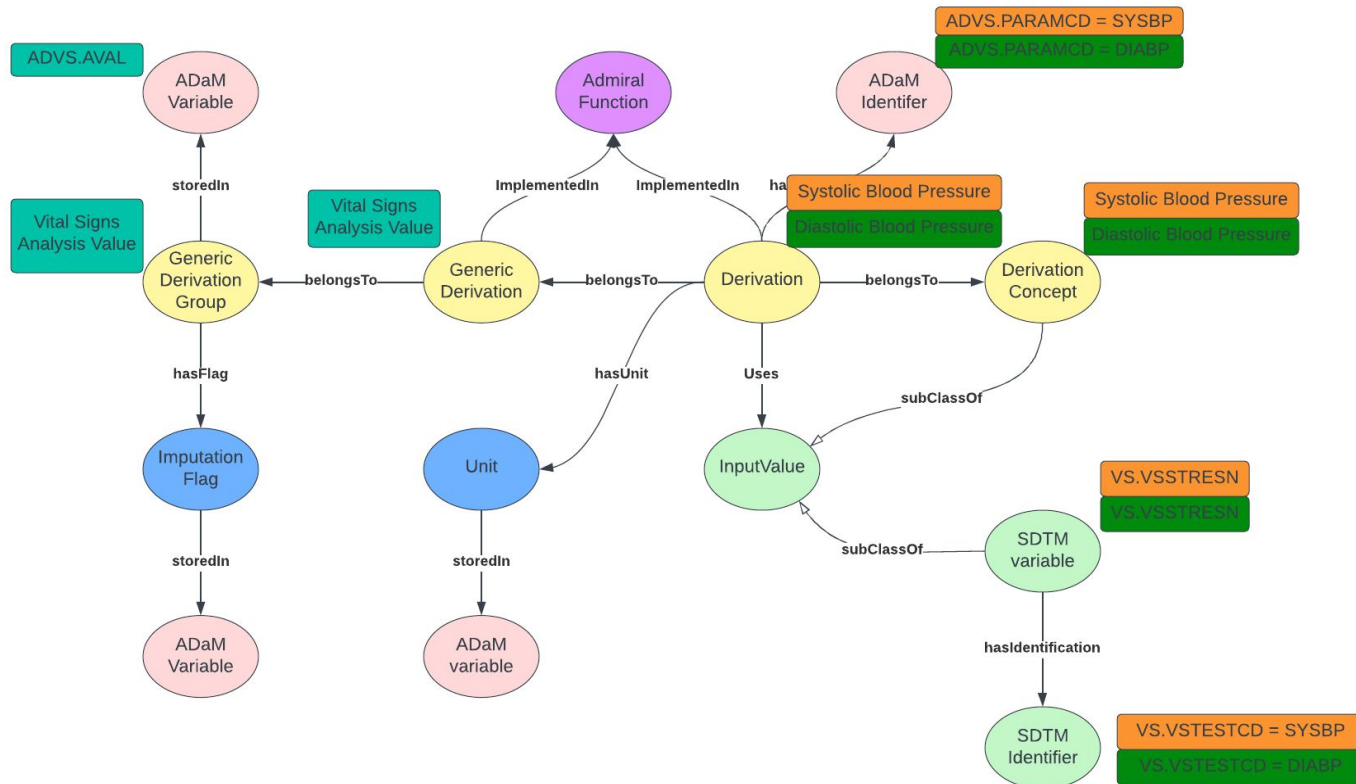


Example - Analysis Age



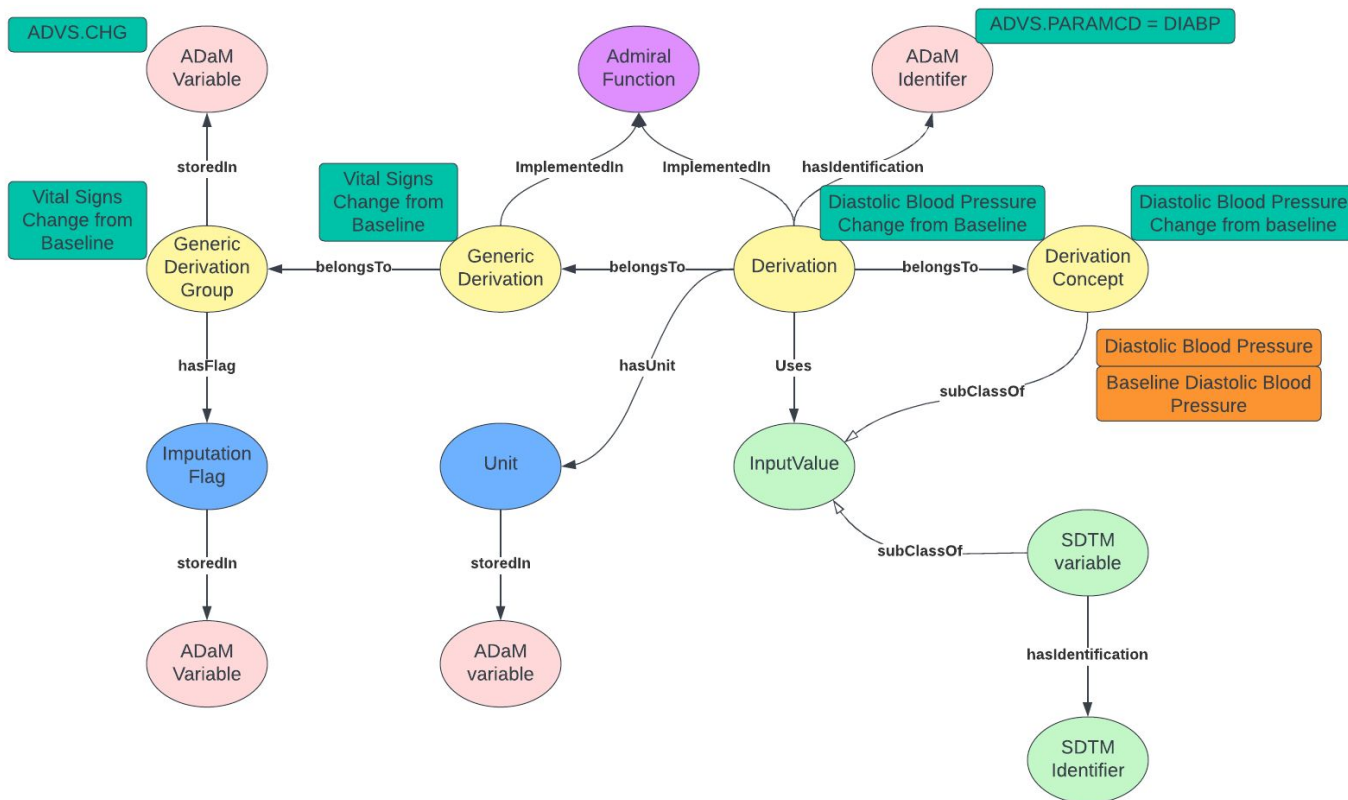
Ontology Development

Example - Vital Sign Analysis Value



Ontology Development

Example - Vital Signs Change from Baseline



Results

Querying the Analysis Dataset Data Standards Ontology



Identify variables needed for a derivation

[← BACK](#) **Query:** *GDSR - Endpoint - PoC - Get all ADaM variables and derivations required to derive and populate a derivation concept*

Description: GDSR - Endpoint - PoC - Please select a derivation concept

Parameters:

derivation concept ⓘ*

Study Day of Death/Censor



RESET

EXECUTE QUERY

EXPORT

Querying the Analysis Dataset Data Standards Ontology



Identify variables needed for a derivation

← BACK Query: GDSR

Results									
20 results.									
dataset	variable	label	programmingSpec	formulaPattern	functionCall	length	controlledTerms		
Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter		
ADRS	PARAM	Parameter	See ADRS Parameters			200			
ADRS	PARAMCD	Parameter Code	See ADRS Parameters			8			
ADRS	AVALC	Analysis Value (C)	See ADRS Parameters			200			
ADRS	ADT	Analysis Date	See ADRS Parameters			8			
ADSL	RANDDT	Date of Randomization	Set to datepart of patient's Randomization date: Start Date/Time of Disposition Event [DS.DSSTDTC...	derive_disposition_dt	derive_disposition_dt(dataset = dm, dataset_ds = ds, new_var = RANDDT, dtc = DSSTDTC, filter = DS...	8			
ADSL	TRTSDTM	Datetime of First Exposure to Treatment	Set to datetime of patient's first exposure observation Start Date/Time of Treatment [EX.EXSTDTC]...	derive_var_trtsdtm	derive_var_trtsdtm(, dataset_ex = ex)	8			
ADSL	TRTSDT	Date of First Exposure to Treatment	Set to the date part of the datetime of Last Exposure to Treatment [ADSL.TRSTDTM]			8			
ADSL	TRTSTMF	Time of First Exposure Imput. Flag	Set to "H" if the entire time is imputed. Set to "M" if minutes and seconds are imputed. Set t...	derive_var_trtsdtm	derive_var_trtsdtm(, dataset_ex = ex)	1	http://rdf.cdisc.org/adam-terminology#C81226		
ADSL	TRTEDTM	Datetime of Last Exposure to Treatment	Set to datetime of patient's last exposure observation End Date/Time of Treatment [EX.EXENDTC] con...	derive_var_trtedtm	derive_var_trtedtm(, dataset_ex = ex)	8			
ADSL	TRTEDT	Date of Last Exposure to Treatment	Set to the date part of the datetime of Last Exposure to Treatment [ADSL.TRTEDTM]			8			
ADSL	TRTETMF	Time of Last Exposure Imput. Flag	Set to "H" if the entire time is imputed. Set to "M" if minutes and seconds are imputed. Set t...	derive_var_trtedtm	derive_var_trtedtm(, dataset_ex = ex)	1	http://rdf.cdisc.org/adam-terminology#C81226		
ADSL	DTHDT	Date of Death	Set to the datepart of Date of Death [ADSL.DTHDTC] converted to numeric date. No imputation of mli...	derive_disposition_dt	derive_vars_dt(, new_vars_prefix = "DTH", dtc=DTHDTC, flag_imputation = FALSE)	8			
ADSL	LSTALVDT	Date Last Known Alive	Set to the last date patient has documented clinical data to show him/her alive, converted to nu...			8			

Description: GDSR - E

Parameters:

derivation concept ⓘ

Study Day of Death/Censor

CUTE QUERY

EXPORT

Querying the Analysis Dataset Data Standards Ontology



Identify parameters needed for a derivation

[← BACK](#) **Query:** *GDSR - Endpoint - PoC - Get all ADaM variables and derivations that vary based on the parameter that are required to derive and populate a derivation concept*

Description: GDSR - Endpoint - PoC - Please select a derivation concept

Parameters:

derivation concept 

Study Day of Death/Censor



RESET

EXECUTE QUERY

EXPORT

Querying the Analysis Dataset Data Standards Ontology

Identify parameters needed for a derivation

← BACK Query: GDSR - Endpoint - PoC - Get all ADaM variables and derivations that vary based on the parameter that are required to derive and populate a derivation concept

Results
8 results. Show columns (10/21) CLEAR FILTERS

Description

Parameter derivation con Study Day

dataset	variable	label	ProgrammingSpec	IdentifierText	IdentifierCode	unit_value	controlledTerms	displayFormat	sasFormat
Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter
ADRS	AVALC	Analysis Value (C)	Set to 'Y' if [ADSL.DTHDT] is not missing Else set to 'N' for all other subjects in ADSL.	Death	DEATH				
ADRS	ADT	Analysis Date	Set to the date of death [ADSL.DTHDT]	Death	DEATH				date9.
ADTTE	AVALU	Analysis Value Unit		Overall Survival	OS	DAYS	http://rdf.cdisc.org/sdtm-terminology#C71620	Unit	
ADTTE	STARTDT	Time-to-Event Origin Date for Subject	Set to Date of First Exposure to Treatment [ADSL.TRTSDT]	Overall Survival	OS				date9.
ADTTE	STARTDT	Time-to-Event Origin Date for Subject	Set to Date of Randomization [ADSL.RANDDT]	Overall Survival	OS				date9.
ADTTE	STARTDT	Time-to-Event Origin Date for Subject	Set to Date of Randomization [ADSL.RANDDT]	Overall Survival	OS				date9.
ADTTE	ADT	Analysis Date	Set to Death date [ADRS.ADT] if there is an observation in ADRS, where [ADRS.PARAMCD]='DEATH' and...	Overall Survival	OS				date9.
ADTTE	ADT	Analysis Date	Set to Death date [ADRS.ADT] if there is an observation in ADRS, where [ADRS.PARAMCD]='DEATH' and...	Overall Survival	OS				date9.

EXPORT

Querying the Analysis Dataset Data Standards Ontology



Identify inputs from SDTM

[← BACK](#) Query: *GDSR - Endpoint - PoC -Get all SDTM variables potentially involved in a derivation concept*

Description: GDSR - PoC (test only) Please select a derivation concept

Parameters:

derivation concept ⓘ

Study Day of Death/Censor



RESET

EXECUTE QUERY

EXPORT

Querying the Analysis Dataset Data Standards Ontology



Identify inputs from SDTM

← BACK

Query: *GDSR - Endpoint - PoC - Get all SDTM variables potentially involved in a derivation concept*

Results

22 results.

Show columns (5/6)

CLEAR FILTERS

Description:

GDSR

Parameters:

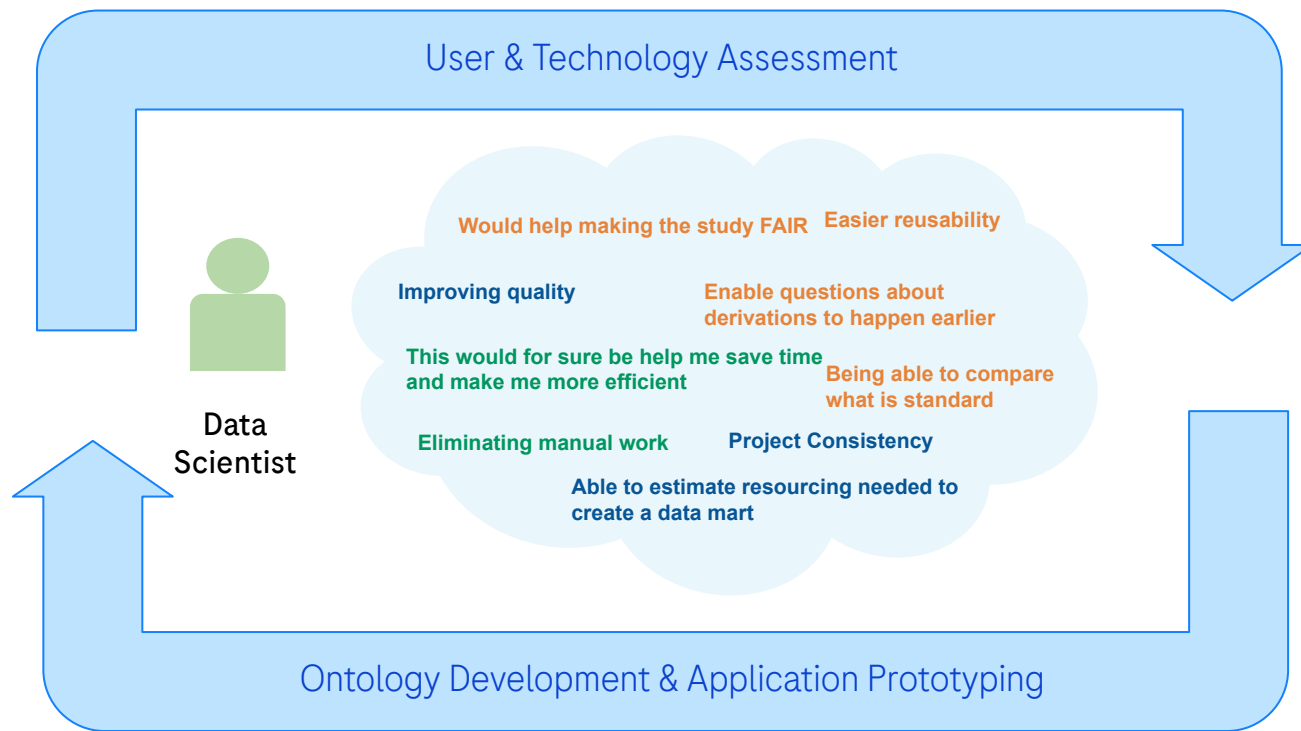
derivation concept ⓘ

Study Day of Death/Censor

Derivation Concept Label	Column URI	SDTM Domain	SDTM Variable	Value
Filter	Filter	Filter	Filter	Filter
Study Day of Death/Censor	http://gdsr.roche.com/cdisc/sdtmig#Column.AE.AESTDTC	AE	AESTDTC	
Study Day of Death/Censor	http://gdsr.roche.com/cdisc/sdtmig#Column.AE.AEENDTC	AE	AEENDTC	
Study Day of Death/Censor	http://gdsr.roche.com/cdisc/sdtmig#Column.LB.LBDTC	LB	LBDTC	
Study Day of Death/Censor	http://gdsr.roche.com/cdisc/sdtmig#Column.SS.SSDTC	SS	SSDTC	
Study Day of Death/Censor	http://gdsr.roche.com/cdisc/sdtmig#Column.RS.RSDTC	RS	RSDTC	
Study Day of Death/Censor	http://gdsr.roche.com/cdisc/sdtmig#Column.LB.LBTESTCD	LB	LBTESTCD	
Study Day of Death/Censor	http://gdsr.roche.com/cdisc/sdtmig#Column.LB.LBSTAT	LB	LBSTAT	NOT DONE
Study Day of Death/Censor	http://gdsr.roche.com/cdisc/sdtmig#Column.SS.SSTESTCD	SS	SSTESTCD	SURVSTAT
Study Day of Death/Censor	http://gdsr.roche.com/cdisc/sdtmig#Column.SS.SSSTRESC	SS	SSSTRESC	ALIVE
Study Day of Death/Censor	http://gdsr.roche.com/cdisc/sdtmig#Column.RS.RSSTRESC	RS	RSSTRESC	SD
Study Day of Death/Censor	http://gdsr.roche.com/cdisc/sdtmig#Column.RS.RSSTRESC	RS	RSSTRESC	CR
Study Day of Death/Censor	http://gdsr.roche.com/cdisc/sdtmig#Column.RS.RSSTRESC	RS	RSSTRESC	NON-CR/NON-PD
Study Day of Death/Censor	http://gdsr.roche.com/cdisc/sdtmig#Column.RS.RSSTRESC	RS	RSSTRESC	PD
Study Day of Death/Censor	http://gdsr.roche.com/cdisc/sdtmig#Column.RS.RSSTRESC	RS	RSSTRESC	PR
Study Day of Death/Censor	http://gdsr.roche.com/cdisc/sdtmig#Column.RS.RSTESTCD	RS	RSTESTCD	OVRLRESP

EXECUTE QUERY

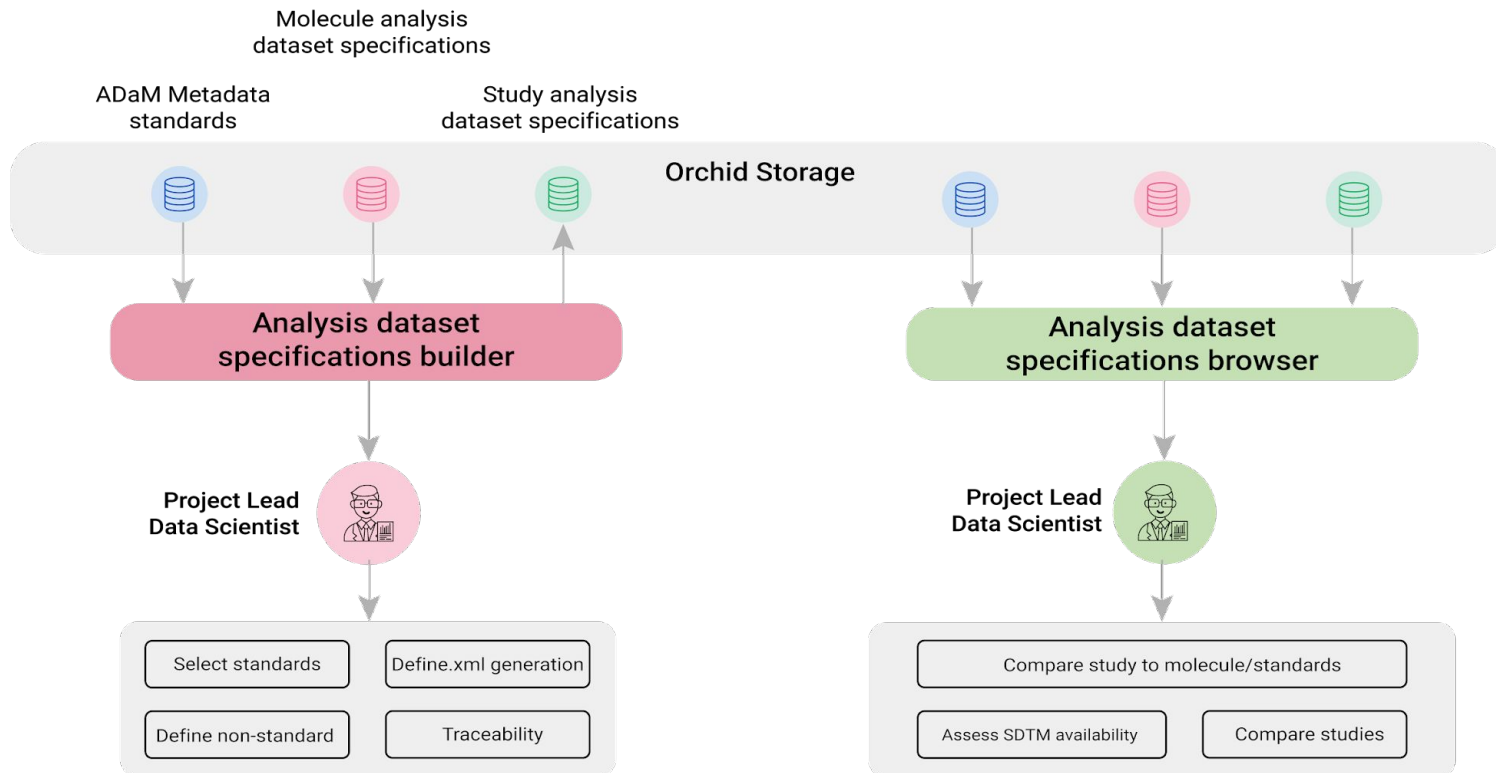
EXPORT



Next Steps and Outlook

Next Steps: Orchid Garden Elements

Application, Connectivity and Graphs



The Team

Product Owner Network

Nelia Lasierra

Chris Price

Robin Koeger

Selena Baset

Extended Core Team

Didier Clement

Michal Pietrusinski

Ivan Robinson

Julian Jager

SMEs

Kathrin Flunkert

Jennifer Lomax

Samia Kabai

Rashad Rasul*

Technology Development

Magdalena Krochmal

Paulina Frischke

Keo Chanthavinout

Data Standards

Cathy Yang

Graphs

Huw Mason

Javier D. Fernández

Project Botany

Eddy Foster

Ross Farrugia

Doing now what patients need next