

Data Modeling in Gen3 Data Commons

Gen3 Community Forum
6 July 2023

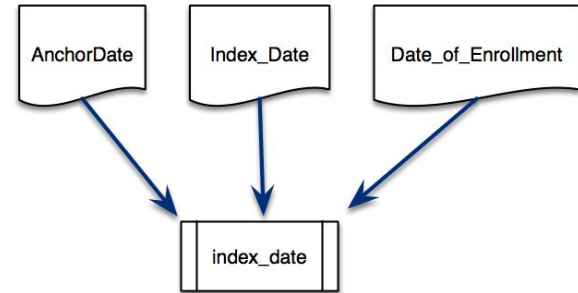
- Introduction to Gen3 Data Models
- Data Commons Presentations
 - **Evolution of the MIDRC Data Model** (Chris Meyer - Center for Translational Data Science, University of Chicago)
 - **Streamlining Gen3 Data Dictionaries: Python Tools and Google Sheets for simple, automated and efficient dictionary development** (Marion Shadbolt - Australian BioCommons)
 - **Spreadsheet-based data ingest with Gen3 dictionary-based validation** (Eirian Perkins - New Zealand eScience Infrastructure (NeSI))
 - **Versioning, migrations, and data release processes in the Pediatric Cancer Data Commons** (Brian Furner - Data for the Common Good, University of Chicago)
- Discussion

Introduction to Gen3 Data Models

Michael Fitzsimons

- What is a data model and a data dictionary?
- The structure of a Gen3 data model
- Tips for creating a Gen3 data model

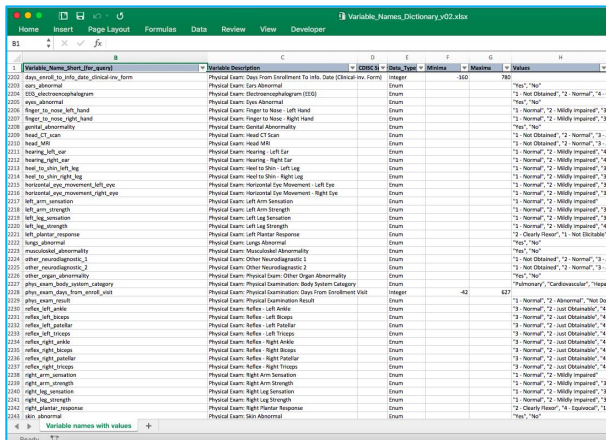
- The data dictionary defines and describes how research datasets are represented in the database and harmonizes/aligns term definitions from different data sources
- Dictionaries get everyone on the same page:
 - Defines nodes and properties used across different but similar projects.
 - Help avoid inconsistencies in data reporting and use across projects.
 - Make data easier to find, subset and analyze by enforcing Data Standards.



Data Dictionary vs Data Model

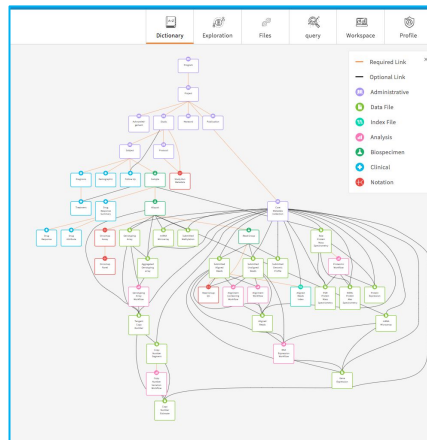
- A **data model** organizes terms from a **data dictionary** and defines how they relate to one another. It is the implementation of a data dictionary and enables gen3 services to submit, index, and query data

Data Dictionary

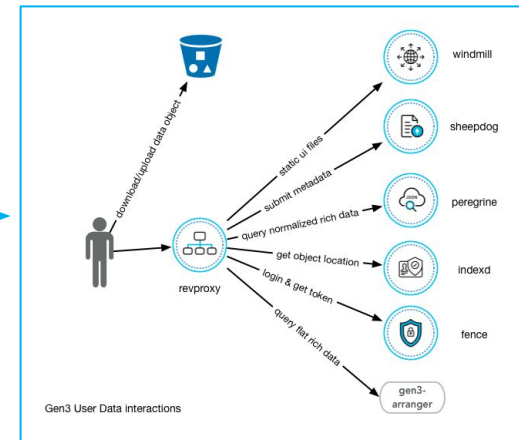


Variable Name	Description	Unit	Value
2200	Physical Exam: Eyes from Endowment To Info, Date (Clinical On-Span)	Integer	365
2201	Physical Exam: Iops Abnormal	Enum	"No", "Yes"
2202	Physical Exam: Electroencephalogram (EEG)	Enum	"1", "Not Obtainable", "2", "Normal", "4", "C"
2203	Physical Exam: Eye Abnormal	Enum	"No", "Yes"
2204	Physical Exam: Finger to Nose - Left Hand	Enum	"1", "Normal", "2", "Mildly Impaired", "3"
2205	Physical Exam: Finger to Nose - Right Hand	Enum	"1", "Normal", "2", "Mildly Impaired", "3"
2206	Physical Exam: Sensible Abnormality	Enum	"No", "Yes"
2207	Physical Exam: Head CT Scan	Enum	"1", "Not Obtainable", "2", "Normal", "3", "A"
2208	Physical Exam: Head MRI	Enum	"1", "Not Obtainable", "2", "Normal", "3", "A"
2209	Physical Exam: Hearing - Right Ear	Enum	"1", "Normal", "2", "Mildly Impaired", "3"
2210	Physical Exam: Hearing - Left Ear	Enum	"1", "Normal", "2", "Mildly Impaired", "3"
2211	Physical Exam: Hearing - Right Ear	Enum	"1", "Normal", "2", "Mildly Impaired", "3"
2212	Physical Exam: Hearing - Left Ear	Enum	"1", "Normal", "2", "Mildly Impaired", "3"
2213	Physical Exam: Hand to Shin - Left Leg	Enum	"1", "Normal", "2", "Mildly Impaired", "3"
2214	Physical Exam: Hand to Shin - Right Leg	Enum	"1", "Normal", "2", "Mildly Impaired", "3"
2215	Physical Exam: Horizontal Eye Movement - Left Eye	Enum	"1", "Normal", "2", "Mildly Impaired", "3"
2216	Physical Exam: Horizontal Eye Movement - Right Eye	Enum	"1", "Normal", "2", "Mildly Impaired", "3"
2217	Physical Exam: Left Arm Sensation	Enum	"1", "Normal", "2", "Mildly Impaired", "3"
2218	Physical Exam: Left Arm Strength	Enum	"1", "Normal", "2", "Mildly Impaired", "3"
2219	Physical Exam: Left Leg Sensation	Enum	"1", "Normal", "2", "Mildly Impaired", "3"
2220	Physical Exam: Left Leg Strength	Enum	"1", "Normal", "2", "Mildly Impaired", "3"
2221	Physical Exam: Left Plantar Response	Enum	"1", "Normal", "2", "Mildly Impaired", "3"
2222	Physical Exam: Left Plantar Response	Enum	"1", "Normal", "2", "Mildly Impaired", "3"
2223	Physical Exam: Left Plantar Response	Enum	"1", "Normal", "2", "Mildly Impaired", "3"
2224	Physical Exam: Left Plantar Response	Enum	"1", "Normal", "2", "Mildly Impaired", "3"
2225	Physical Exam: Left Plantar Response	Enum	"1", "Normal", "2", "Mildly Impaired", "3"
2226	Physical Exam: Left Plantar Response	Enum	"1", "Normal", "2", "Mildly Impaired", "3"
2227	Physical Exam: Left Plantar Response	Enum	"1", "Normal", "2", "Mildly Impaired", "3"
2228	Physical Exam: Left Plantar Response	Enum	"1", "Normal", "2", "Mildly Impaired", "3"
2229	Physical Exam: Left Plantar Response	Enum	"1", "Normal", "2", "Mildly Impaired", "3"
2230	Physical Exam: Left Plantar Response	Enum	"1", "Normal", "2", "Mildly Impaired", "3"
2231	Physical Exam: Left Plantar Response	Enum	"1", "Normal", "2", "Mildly Impaired", "3"
2232	Physical Exam: Left Plantar Response	Enum	"1", "Normal", "2", "Mildly Impaired", "3"
2233	Physical Exam: Left Plantar Response	Enum	"1", "Normal", "2", "Mildly Impaired", "3"
2234	Physical Exam: Left Plantar Response	Enum	"1", "Normal", "2", "Mildly Impaired", "3"
2235	Physical Exam: Left Plantar Response	Enum	"1", "Normal", "2", "Mildly Impaired", "3"
2236	Physical Exam: Left Plantar Response	Enum	"1", "Normal", "2", "Mildly Impaired", "3"
2237	Physical Exam: Left Plantar Response	Enum	"1", "Normal", "2", "Mildly Impaired", "3"
2238	Physical Exam: Left Plantar Response	Enum	"1", "Normal", "2", "Mildly Impaired", "3"
2239	Physical Exam: Left Plantar Response	Enum	"1", "Normal", "2", "Mildly Impaired", "3"
2240	Physical Exam: Left Plantar Response	Enum	"1", "Normal", "2", "Mildly Impaired", "3"
2241	Physical Exam: Left Plantar Response	Enum	"1", "Normal", "2", "Mildly Impaired", "3"
2242	Physical Exam: Left Plantar Response	Enum	"1", "Normal", "2", "Mildly Impaired", "3"
2243	Physical Exam: Left Plantar Response	Enum	"1", "Normal", "2", "Mildly Impaired", "3"
2244	Physical Exam: Left Plantar Response	Enum	"1", "Normal", "2", "Mildly Impaired", "3"
2245	Physical Exam: Left Plantar Response	Enum	"1", "Normal", "2", "Mildly Impaired", "3"
2246	Physical Exam: Left Plantar Response	Enum	"1", "Normal", "2", "Mildly Impaired", "3"
2247	Physical Exam: Left Plantar Response	Enum	"1", "Normal", "2", "Mildly Impaired", "3"
2248	Physical Exam: Left Plantar Response	Enum	"1", "Normal", "2", "Mildly Impaired", "3"
2249	Physical Exam: Left Plantar Response	Enum	"1", "Normal", "2", "Mildly Impaired", "3"
2250	Physical Exam: Left Plantar Response	Enum	"1", "Normal", "2", "Mildly Impaired", "3"

Data Model



Gen3 Services



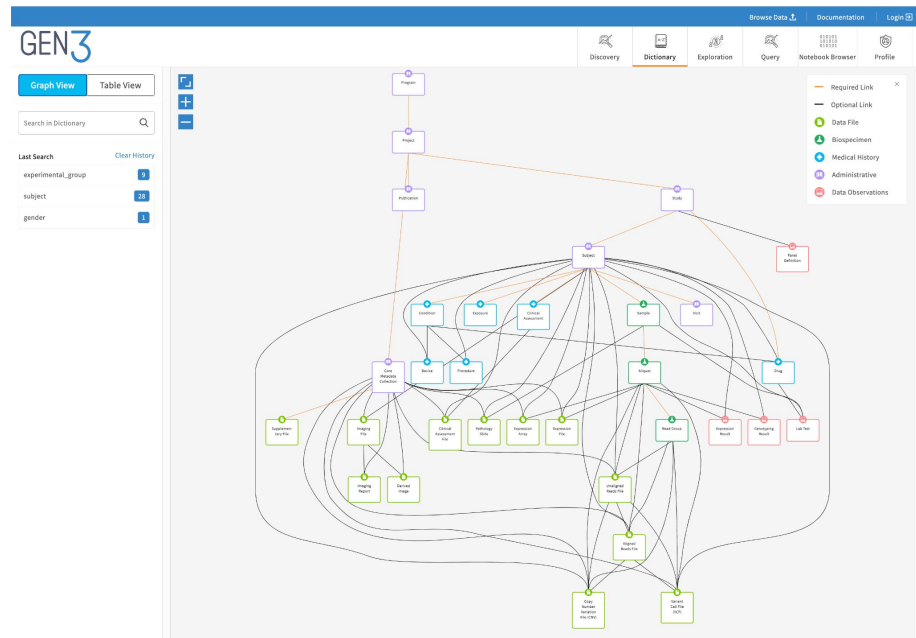
Structure of a Gen3 Data Model

- The Gen3 Data Model is a graph-like relational model consisting of interrelated **nodes** that store certain related **properties**.

The screenshot shows the Gen3 Dictionary interface for the 'Visit' node. The top navigation bar includes 'Browse Data', 'Documentation', and 'Login'. Below the navigation, there are tabs for 'Graph View' and 'Table View'. A search bar is present on the left. The main content area displays the 'Visit' node definition, including a description and a table of properties.

Visit has 19 properties.

Property	Type	Required	Description	Term
type	string	Required	The node_id of the node in the data model, the name of the node used in queries and API requests (e.g., "aligned_reads_file" for the "Aligned Reads File" node).	
submitter_id	string	Required	A human-readable, unique identifier for a record in the metadata database. It can be used in place of the UUID for identifying or recalling a record (e.g., in data queries or uploads/reports).	
subjects	array object	Required	No Description	
days_to_visit	integer	No	The number of days between the case Index Date and the date of the visit, call or other interaction.	
days_to_visit_end	integer	No	The number of days from the index date to the date the patient's visit ended.	
days_to_visit_start	integer	No	The number of days from the index date to the date the patient's visit began.	
months_between_visits	integer	No	The number of months between the case Index Date and the date of the visit, call or other interaction	
subject_ids	array	No	A list of one or more subject submitter_ids associated with this data.	
visit_duration	integer	No	The number of days from the start date of the visit to the end date of the visit.	
visit_epoch	integer	No	If a single visit is sub-divided into multiple time-points or epochs, or if an unscheduled visit takes place between two integer-labeled visits, specify the order of such sub-divisions, timepoints, or unscheduled visits here. For example, specify 'visit_number' of '1' and 'visit_epoch' of '2' for 'Visit 1.2'; if there is no sub-division of a single visit into multiple timepoints or unscheduled visits between numbered visits, submit only 'visit_number'.	
visit_id	string	No	The submitter_id of a subject's visit in the study.	
visit_label	Annual Clinical • Baseline • Biopsy • CTCC Only • Consent • Electronic Health Records • Endpoint • Final • Follow-up • Genetic Testing • Imaging • LIG • Other (specify) • Pre-screening	No	The reason for or context of the visit, call or other interaction. For a generic visit, specify "Visit".	



Structure of a Gen3 Data Model

- The data model is a JSON created from schemas in the YAML format.
- Each node is defined in a single schema.
- The schema contains the following:
 - A node id used for data query/submission.
 - A category used to group nodes conceptually.
 - A description which describes the node's contents
 - List of links defining relationship to other nodes.
 - List of required properties.
 - List of properties.

```
demographic.yaml
1 $schema: "http://json-schema.org/draft-04/schema#"
2 ~
3 id: "demographic"
4 title: Demographic
5 type: object
6 namespace: https://nci-crdc-demo.datacommons.io/
7 category: clinical
8 program: '*'
9 project: '*'
10 description: >
11   Data for the characterization of the patient by means of
12 additionalProperties: false
13 submittable: true
14 validators: null
15 ~
16 systemProperties:
17   - id
18   - project_id
19   - state
20   - created_datetime
21   - updated_datetime
22 ~
23 links:
24   - name: subjects
25     backref: demographics
26     label: describes
27     target_type: subject
28     multiplicity: one_to_one
29     required: true
```


- Collect use cases for the new data commons
 - Not all individual data elements need to be represented in the data model.
 - Some data should simply be stored in data files.
 - Which data elements are represented in the data model as properties depends on how users will query the data.
 - Examples:
 - Clinical properties, e.g., in diagnosis and demographic nodes, can be used to select subject cohorts
 - Biospecimen properties, e.g., in sample, aliquot, or read_group nodes, like collection or processing properties can be used to subset data files
 - Data_file properties can be used to filter file types and formats

Tips for Creating a Gen3 Data Model

- Include references to external vocabularies
 - In order to facilitate data standardization and harmonization, pointers can be used to connect terms to external controlled vocabularies
 - Some examples used by Gen3 commons include NCIt and LOINC

BloodPAC Data Commons

Demographic: Data for the characterization of the patient by means of segmenting the population (e.g., characterization by age, sex, or race).

Demographic has 11 properties.

Property	Type	Required	Description	Term
type	• string	★ Required	No Description	
submitter_id	• string	★ Required	A project-specific identifier for a node. This property is the calling card/nickname/alias for a unit of submission. It can be used in place of the UUID for identifying or recalling a node.	
cases	• array • object	★ Required	No Description	
cause_of_death	• Cancer Related • Not Cancer Related • Unknown	No	Text term to identify the cause of patient death with respect to cancer.	
days_to_birth	• integer	No	The number of days between the index date and the date of patient birth. If the number of days is greater than 32872 (89 years), then please use 'days_to_birth_g89'.	
days_to_birth_g89	• Yes • No	No	Indicate if the number of days between the index date and the date of patient birth is greater than 32872 (89 years).	
days_to_death	• integer	No	The number of days between the index date and the date of patient death.	
ethnicity	• Hispanic or Latino • Not Hispanic or Latino • Unknown	No	An individual's self-described social and cultural grouping, specifically whether an individual describes themselves as Hispanic or Latino. The provided values are based on the categories defined by the U.S. Office of Management and Business and used by the U.S. Census Bureau.	2192217
gender	• Female • Male • Unknown • Unspecified	No	Text designations that identify gender. Gender is described as the assemblage of properties that distinguish people on the basis of their societal roles. [Explanatory Comment 1: Identification of gender is based upon self-report and may come from a form, questionnaire, interview, etc.]	2200604
race	• White • American Indian or Alaska Native • Black or African American • Asian • Native Hawaiian or Other Pacific Islander • Other	No	An arbitrary classification of a taxonomic group that is a division of a species. It usually arises as a consequence of geographical isolation within a species and is characterized by shared heredity, physical attributes and behavior, and in the case of humans, by common history, nationality, or geographic distribution. The provided values are based on the categories defined by the U.S. Office of Management and Business and used by the U.S. Census Bureau.	2192199

Create New Data Model: External vocabularies

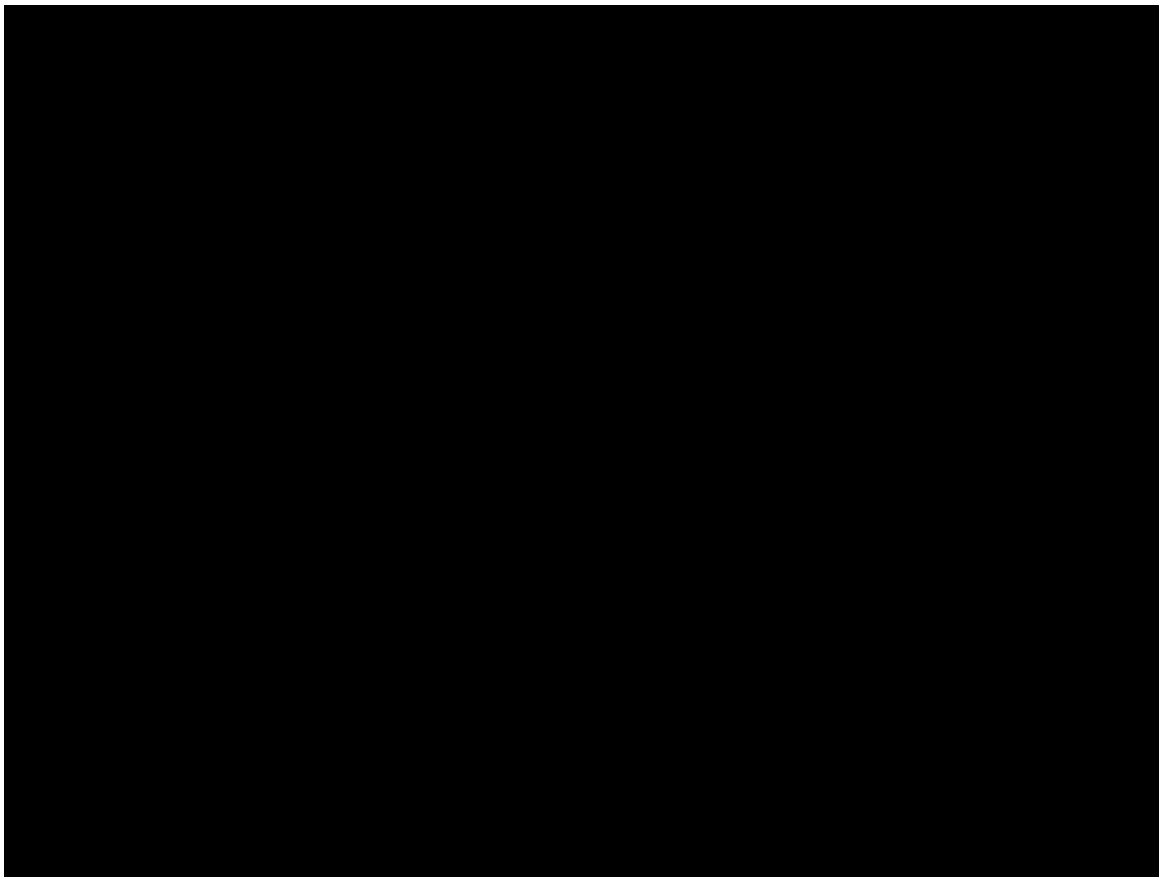
- use termDef for *node* and *properties*
- use enumDef for *enumerated values*

```
56
57 phenotype:
58   description: "Name given to Phenotype by contributor"
59   enum:
60     - "abnormal aorta"
61     - "abnormal aortic valve"
62     - "very dry skin"
63   termDef:
64     - term: phenotype
65     - source: NCI Thesaurus
66     - term_id: C16977
67     - term_version: 18.10.e (Release date:2018-10-29)
68   enumDef:
69     - enumeration: abnormal aorta
70       source: hp
71       term_id: HP:0001679
72       version_date: 2019-02-12
73     - enumeration: abnormal aortic valve
74       source: hp
75       term_id: HP:0001646
76       version_date: 2019-02-12
77     - enumeration: very dry skin
78       source: hp
79       term_id: HP:0000958
80       version_date: 2019-02-12
81
```

Evolution of the MIDRC Data Model

Chris Meyer

Evolution of the MIDRC Data Model

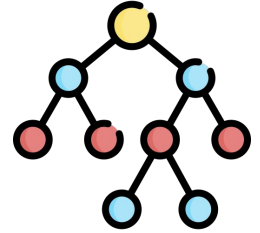
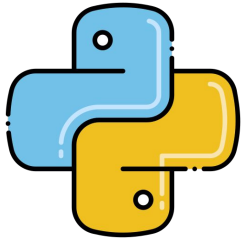


Streamlining Gen3 Data Dictionaries: Python Tools and Google Sheets for simple, automated and efficient dictionary development

Marion Shadbolt

Streamlining Gen3 Data Dictionaries:

Python Tools and Google Sheets for simple, automated and efficient dictionary development



Gen3 Data Modelling User Forum - Thursday, July 6 / Friday, July 7 2023

Uwe Winter & Marion Shadbolt

Outline

- Context and Overview
- Harmonizing Objects, Fields, Data types
- Compiling the Dictionary
- Visualizing the Dictionary structure
- Visualizing the portal with content

Australian Cardiovascular disease Data Commons



ACDC

Data Governance

Legal Ethics Policy Trust

Multi-omics cohort data

Genomics
Lipidomics
Metabolomics
Demographic
Clinical
Imaging

Data Platform

Harmonise

Store

Discover

Share

Analysis Platform

Integrate

Analyse

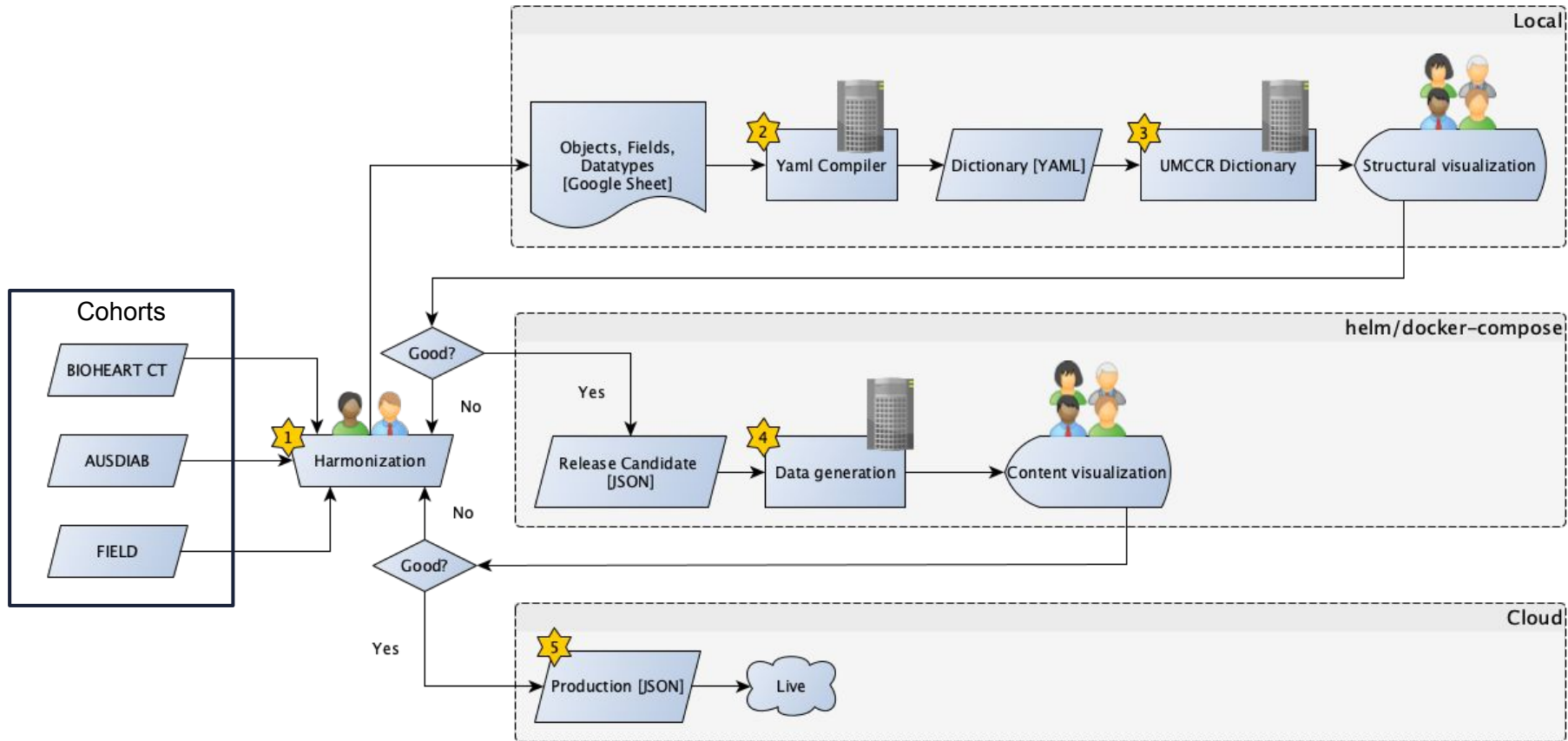
Collaborate

Publish

Translational Impact

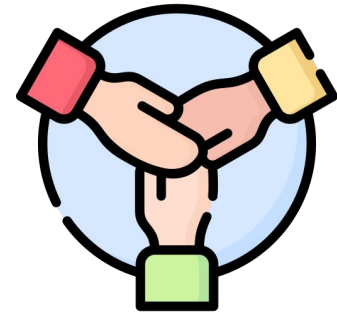
Enhanced understanding of cardiovascular risk, leading to better health outcomes

Data dictionary development workflow



Harmonizing Objects, Fields and Datatypes

- Initially harmonise 20-30 variables
- Engage with data custodians
- Use [BioData Catalyst](#) for structure
- Align to standards
 - ontologies
 - standard identifiers
(e.g. Human Metabolome Database (HMDB) ID)



Harmonizing Objects, Fields and Datatypes

	NHLBI - TOPMed Program - >85 Studies	NHLBI - BioData Catalyst	U.S. National Cardiovascular Research Infrastructure Project	Australian Institute of Health and Welfare - Cardiovascular Diseases	AusDiab	BioHEART	FIELD	HARMONISED VARIABLE	No. of Variables 36	Working Group Notes
Variable Name	bp_diastolic_1	bp_diastolic	diastolicBloodPressure	Blood pressure - diastolic	DBP	dbp	b_diastolic	bp_diastolic	1	Potentially include background notes on measurement
Description	Resting diastolic blood pressure from the upper arm in a clinical setting.	Resting diastolic blood pressure from the upper arm in a clinical setting.	Diastolic blood pressure	The person's diastolic blood pressure	Diastolic blood pressure. Mean diastolic blood pressure (av of closest 2 if 3 measures taken, or first 2 if close enough)	Diastolic blood pressure	Baseline Diastolic Blood Pressure (mmHg) (average of V1, V2 and V3)	Diastolic blood pressure at baseline		
Measured/self-reported	Measured	Measured	Measured	measured	measured	measured	measured	Measured		
Data set	Blood pressure	blood pressure test	Physical exam					Clinical\Blood Pressure		
Type		decimal	Integer	Number	continuous, integer	continuous, int	Numeric	Continuous, integer		
Units	mmHg	mmHg	mmHg	mmHg	mmHg	mmHg	mmHg	mmHg		
Values										
How question is asked?							Registration/Screening Visit 1 (-16 weeks) Run-In Phase I Visit 2 (-12 weeks) Run-In Phase II Visit 3 (-6 weeks)			

Compiling: Sheets->YAML

Objects

ID	TITLE	CATEGORY	DESCRIPTION
project	Project	administrative	The study the data is con
pub	Publication	administrative	Publication for a project
ack	Acknowledgement	administrative	Acknowledgement of con

Links

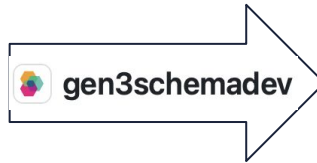
SCHEMA	NAME	PARENT	BACKREF	LABEL	MU
aligned_reads_file	samples	sample	aligned_reads_files	data_from	ma
sub	aligned_reads_file	core_metadata_collection	aligned_reads_files	data_from	ma
lab	aligned_reads_file	aligned_reads_file	aligned_reads_index_file	describes	one

Variables

VARIABLE_NAME	OBJECT	REQUIRED	TYPE	DESCRIPTION
contact_type	acknowledgement	TRUE	enum_role	The type of contact c
orcid	acknowledgement	FALSE	string	The ORCID number
ackno	acknowledgement	TRUE	string	Name of the individu

Enums

type_name	enum	enum_definition	source	term_id
enum_diabetes	IGT	IGT	hpo	HP:0040270
enum_diabetes	KDM	KDM		
enum_diabetes	IFG	IFG		
enum_diabetes	NDM	NDM	SNOMED	870528001
enum_diabetes	NGT	NGT	SNOMED	166926006
enum_du	GRU	GRU	duo	DUO:0000042
enum_du	HMB	HMB	duo	DUO:0000006
enum_du	DS	DS	duo	DUO:0000007
enum_du	NPUNCU	NPUNCU	duo	DUO:0000018
enum_du	IRB	IRB	duo	DUO:0000021
enum_du	US	US	duo	DUO:0000026

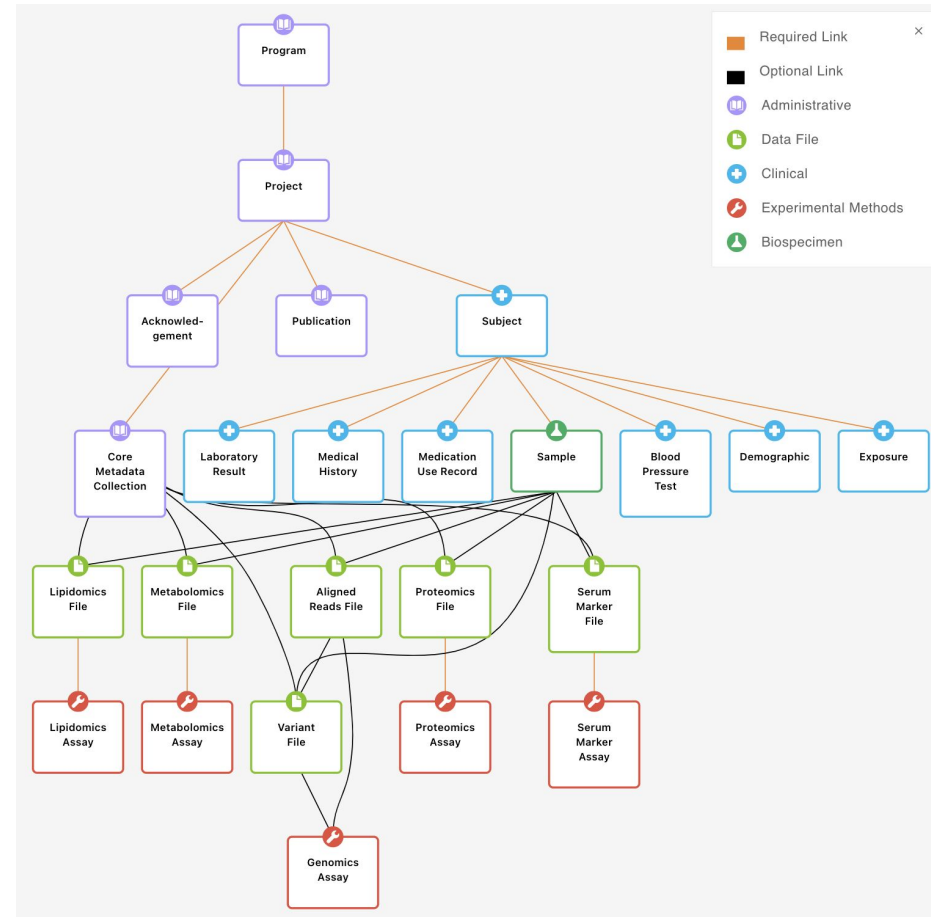


definitions.yaml
_settings.yaml
_terms.yaml
acknowledgement.yaml
aligned_reads_file.yaml
blood_pressure_test.yaml
core_metadata_collection.yaml
demographic.yaml
exposure.yaml
lab_result.yaml
lipidomics_assay.yaml
lipidomics_file.yaml
medical_history.yaml
medication.yaml
program.yaml
project.yaml
publication.yaml
sample.yaml
sequencing_file.yaml
subject.yaml
variant_file.yaml

<https://github.com/AustralianBioCommons/gen3schemadev/tree/main/gen3schemadev>

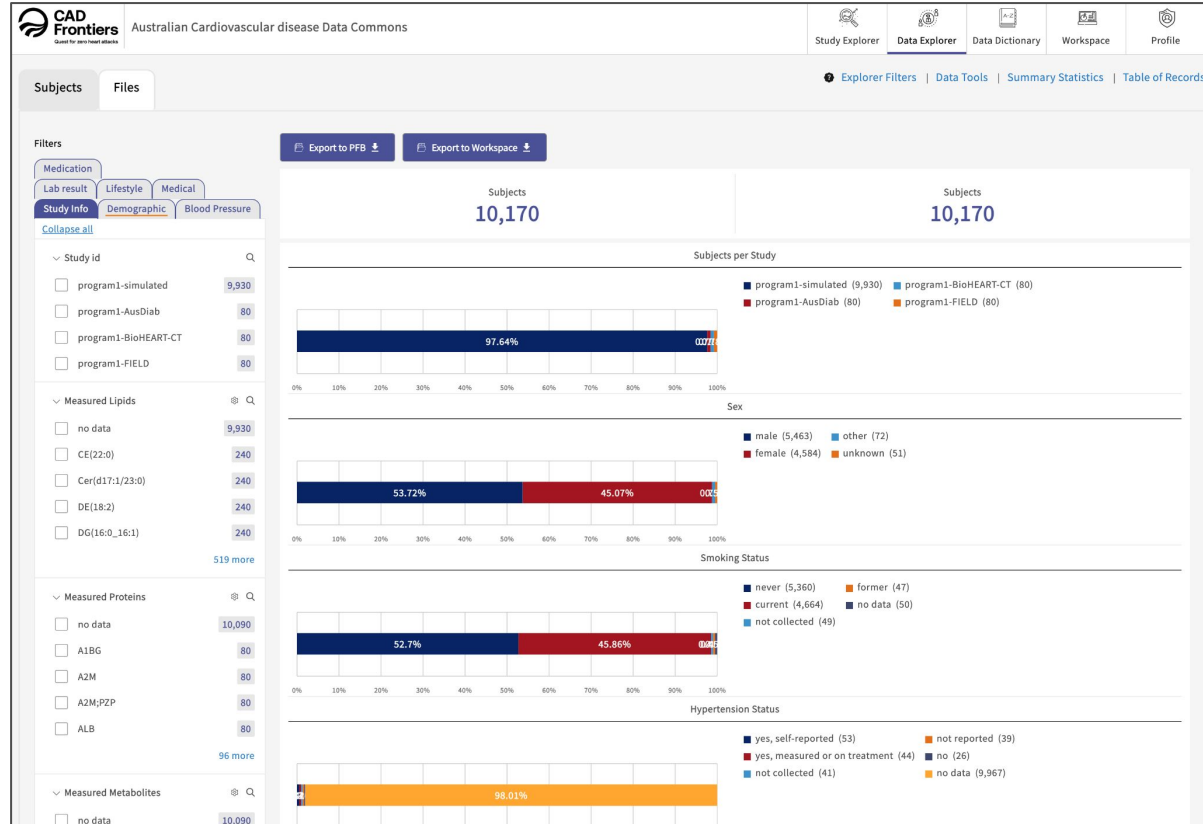
Visualizing the structure

- [UMCCR-dictionary tool](#) for testing, validation, compiling to JSON and visualisation
- Load in local install for review



Visualizing the content

- Adjust ETL, tabs, filters, gitops
- Data linkage
- Get user feedback



What's next?

- Continuing the project with funding from the [MRFF](#) (federal government), Bioplatforms Australia and support from partners
- Aiming to get data from 18 cohorts, ~400,000 individuals into the platform
- Fun times ahead with the wrangling of data into the platform...

Acknowledgements



AusDiab

BioHEART-CT

FIELD

CAD Collaboration

Technical Partner

Peter Miekle

Gemma Figtree

Tony Keech

Tony Willis

Jess Holliday

Dianna Magliano

Michael Gray

Rebecca Mister

Catherine Shang

Marion Shadbolt

Corey Giles

Tung Nguyen

Liping Li

Kerry Doyle

Uwe Winter

Guy Krippner

Jean Yang

Talia Palacios

Steven Manos

Jeff Christiansen

Flaticon Icon attributions:

Slide 1: [Python file icon](#) created by Flat Icons, [Google sheets](#), [Node](#) and [Dictionary](#) icons created by Freepik, [Output icon](#) created by Parzival' 1997




Nuwan Goonasekera

Bernie Pope

Thanks!

 email us at:
technical stuff: uwe@biocommons.org.au
dictionary stuff: marion@biocommons.org.au

 repos:
UMCCR dictionary tool: <https://github.com/umccr/umccr-dictionary>
Schema mapping and compiler tools:
<https://github.com/AustralianBioCommons/gen3schemadev>

Spreadsheet-based data ingest with Gen3 dictionary-based validation

Eirian Perkins

- Aotearoa Genomic Data Repository (AGDR) project
- A Treaty-compliant data archive for New Zealand's taonga species
- Built in partnership with Genomics-Aotearoa



<https://www.youtube.com/watch?v=IQw3OjQI-NM>

Use Case

- Users more familiar with spreadsheet-based metadata entry
- Maintain a familiar experience
 - Example: Sequence Read Archive

* How do you want to enter your data?

Use built-in editor

Upload a file

For more detailed help with SRA submission please read the [SRA Submission Wizard Help](#).

* Sample name	* Library ID	* Title	* Library strategy	* Library source
1 mym1	mym1	mice 1	RNA-Seq	TRANSCRIPTOMIC
2 mym2	mym2	mice 2	RNA-Seq	TRANSCRIPTOMIC
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				
16				
17				
18				

Use Case

- Example:
Geome

The screenshot shows a spreadsheet template with the following content:

2	
3	Identification and development of DNA barcodes on lobster (<i>Panulirus</i> spp.)
4	Template generated on July 06, 2023
5	Person(s) responsible for data entry []
6	
7	Events, Samples, Tissues Tabs
8	Please fill out each field in the Events", "Samples", "Tissues tabs as completely as possible. Fields in red are required (data cannot be uploaded to the database without these fields). Required fields are usually placed towards the beginning of the template. Some fields have a controlled vocabulary associated with them in the "Lists" tab and are provided as data validation in the project. If you have more than one entry to a field (i.e. a list of publications), please delimit your list with pipes (). Also please make sure that there are no newline characters (=carriage returns) in any of your metadata.
9	"Samples", "Tissues tabs may be re-arranged in any order so long as you don't change the field names.
10	
11	Events_Fields, Samples_Fields, Tissues_Fields Tabs
12	This tab contains column names, associated URIs and definitions for each column.
13	
14	Lists Tab
15	This tab contains controlled vocabulary lists for certain fields. DO NOT EDIT this sheet!
16	
17	Additional Instructions
18	If additional fields are needed to capture all data collected for a project, refer to the Geome Workbench Template Generator (https://geome-db.org/workbench/template) for your project. Additional fields may be added to Samples as long as the data conform to the listed definition and data type.
19	
20	
21	
22	
23	
24	
25	
26	
27	
28	

At the bottom of the spreadsheet, there is a navigation bar with the following tabs: **Instructions**, Events, Samples, Tissues, Events_Fields, Samples_Fields, Tissues_Fields, Lists, and a plus sign (+).

- Metadata preparation from researchers can take a long time (~2 weeks)
- Users want live feedback and validation
 - Geome example

Validation results on Events worksheet, for entity: Event.

1 or more errors found. Must fix to continue. Click each message for details

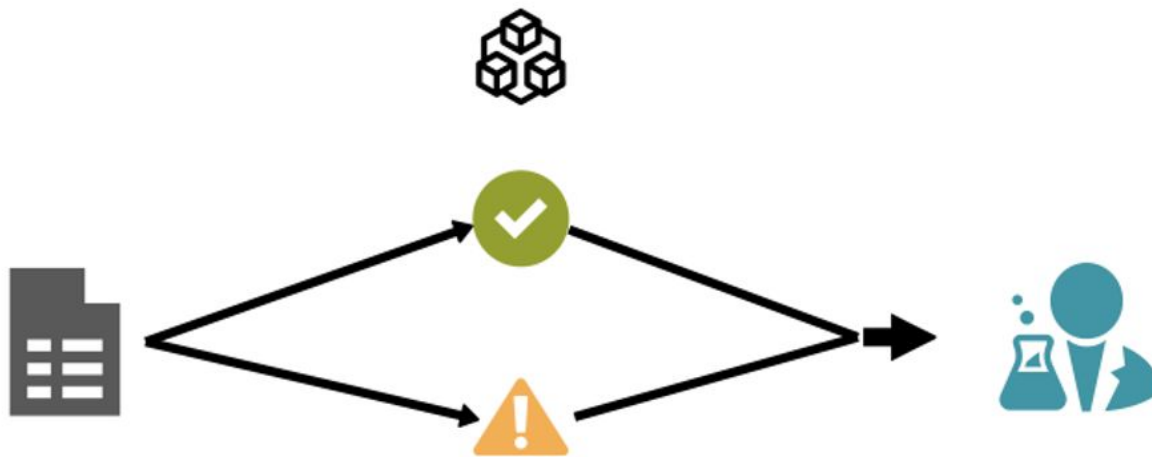
[Error: Missing column\(s\)](#)

"country" has a missing cell value

"locality" has a missing cell value

"yearCollected" has a missing cell value

1. Accept metadata from a spreadsheet template
2. Validate spreadsheet against an arbitrary Gen3 data dictionary
3. Provide feedback to user



1. Metadata ingest template can be manually generated for a particular data repository



Copy of AGDR Metadata Template - 202208

File Edit View Insert Format Data Tools Extensions Help

100% Arial 14

Project Information

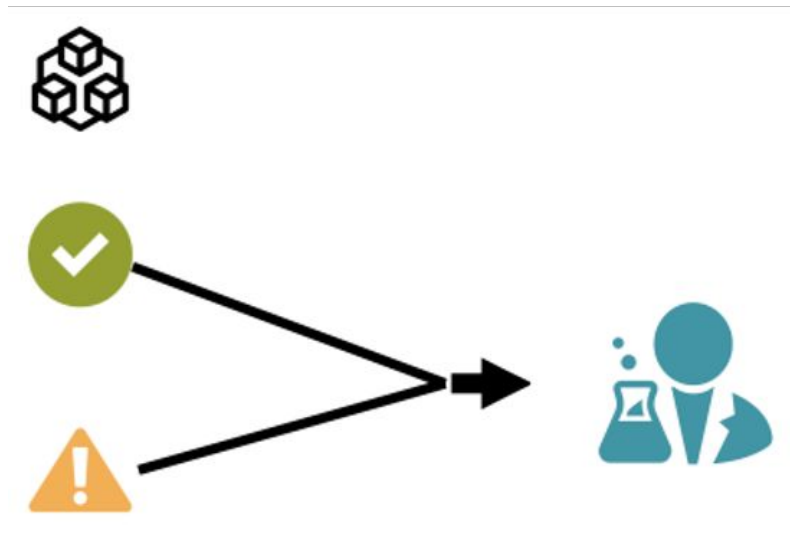
	A	B	C	D	E
1	Project Information				
2	<i>The research project you have been working on. If you have already submitted a Google Forms, you can skip to the next tab, but please read the instructions below first.</i>				
3					
4	Field	name	date_collected	details	investigator_affiliation
5	Required field?	Required	Required		Required
6	Description	Name of the project	The date or date range in which the project data was collected.	More detailed description of the project.	The investigator's affiliation with respect to a research institution.
7	Example input		e.g. 1997-2000	A couple of paragraphs describing the project.	e.g. School of Biomedical Sciences, University of Otago
8	Your Input				
9					
10	Instructions and tips				
11					
12	Please fill out this form to submit your data/metadata into Aotearoa Genomic Data Repository.				
13	This is a template, so please make a copy of this spreadsheet before submitting your input. To make a copy, press 'File' on the menu and 'Make a copy'.				
14	Once you have made a copy of this document, please fill in all the fields as much as you can under 'Your input'. Please note that there are multiple tabs which you can access via the buttons at the bottom.				
15	You can press Alt+Enter for multiline answers if needed.				
16	Once you have completed filling in the details, please remember to share your copied spreadsheet with us with claire.rye@nesi.org.nz , jun.huh@nesi.org.nz , and eirian.perkins@nesi.org.nz ; and NeSI staff will help enter these data into the system.				
17	Please fill in all the 'Required' field and as many optional fields as possible. The required fields are highlighted in blue.				
18					
19	Please feel free to contact us at gasupport@nesi.org.nz for any help.				
20					
21					
22					
23					
24					
25					
26					
27					

1. “submitter_id” is renamed so that it is clearer to users

	A	B	C	
1	Experiments			
2	<i>Experiments done within the project. Please feel free to enter multiple entries by using columns to the right</i>			
3				
4	Field	name or ID	associated_experiment	data_description
5	Required field?	Required	Optional	Optional
6	Description	A unique name/ID for the experiment.	The name/IDs for any experiment with which this experiment is associated, paired, or matched. Comma separated.	Brief description of experiment.
7	Example input	MYEXPERIMENT0001	MYEXPERIMENT0002	FASTQ files of 30 s
8	Your input			
9	<i>Add more rows as needed</i>			
10				
11	Biosamples			
12				
13	<i>Based on the type, please provide more details as seen below. The definitions here have been taken from NCBI (see: https://submit.ncbi.nlm.nih.gov/bios)</i>			
14	Type: Organism			
15				
16	Field	sample_id	experiments	submitted_to_ins
17	Required field?	Required	Required	Required
18	Description	Sample ID is a unique identifier that you choose for the sample. It can have any format, but we suggest that you make it concise, unique, and consistent within your lab. Every Sample ID from a single Submitter must be unique. This will be used in the next tab to link the files to the samples.	List of experiment names/IDs (from above) that this biosample is associated with.	from MixS: Depend with next generation (small-scale) sequencing (Sequence Read Archive, the classical Webi and DDBJ).
19	Example input	MYSAMPLE0001	MYEXPERIMENT0001	false
20	Your input			



2. A data dictionary contains all information necessary to validate draft spreadsheets



1. Parse spreadsheet metadata

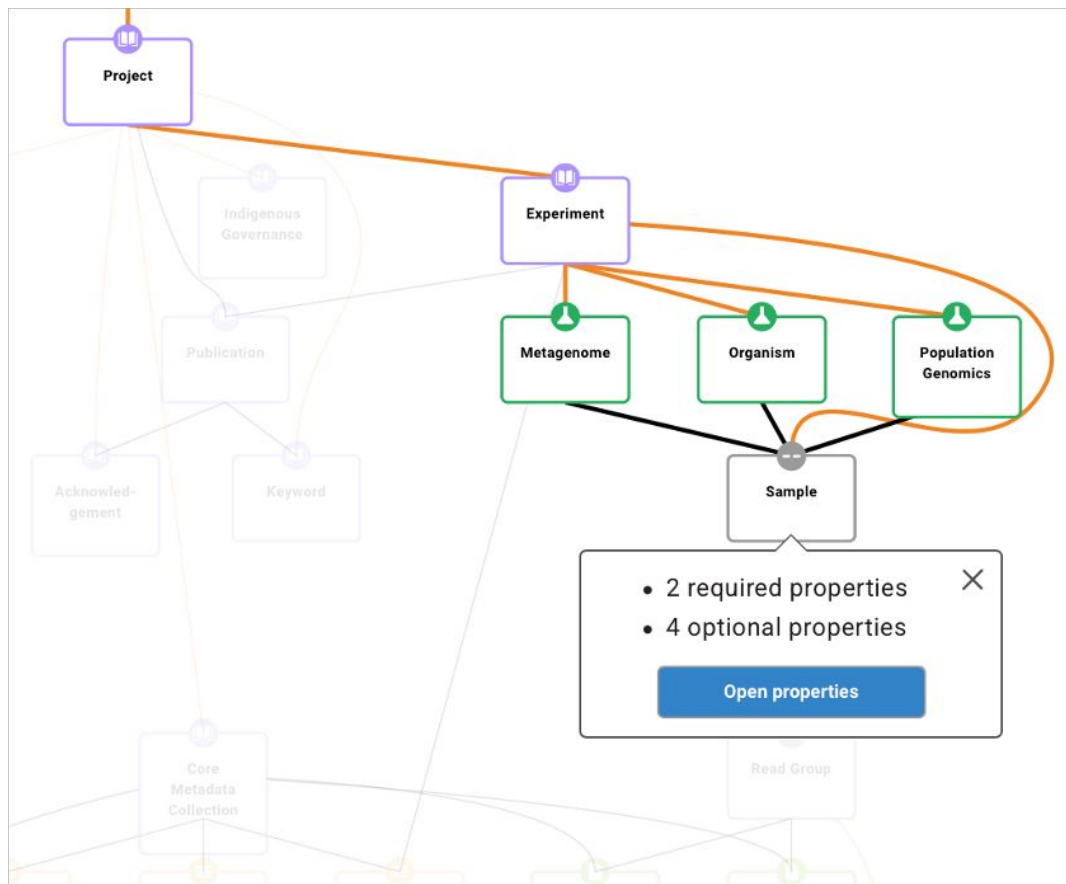
2. Parse data dictionary (JSON schema)

3. Perform rule application on properties and nodes for validation



Data Dictionary Structure

- Main components:
 - Nodes and their properties



- Main components:
 - Nodes and their properties
 - Definitions
 - Can refer to other definitions
 - Can refer to terms

```
1 id: _definitions
2
3 UUID:
4   term:
5     $ref: "_terms.yaml#/UUID"
6   type: string
7   pattern: "^[a-fA-F0-9]{8}-[a-fA-F0-9]{4}-[a-fA-F0-9]{4}-[a-fA-F0-9]{4}-[a-fA-F0-9]{4}$"
8
9 email:
10  term:
11    description: an email address
12  type: string
13  pattern: "[a-zA-Z0-9._]+@[a-zA-Z0-9]+.[a-zA-Z0-9]+"
14
15 parent_uuids:
16  type: array
17  minItems: 1
18  items:
19    $ref: "#/UUID"
20  uniqueItems: true
21
22 foreign_key_project:
23  type: object
24  # Allow true here because we can have other unique keys defined on
25  # a target type
26  additionalProperties: true
27  #Can either use 'id' which are Gen3 IDs (UUID) or 'code'
28  #which is the user defined ID for project
```

- Main components:
 - Nodes and their properties
 - Definitions
 - Terms

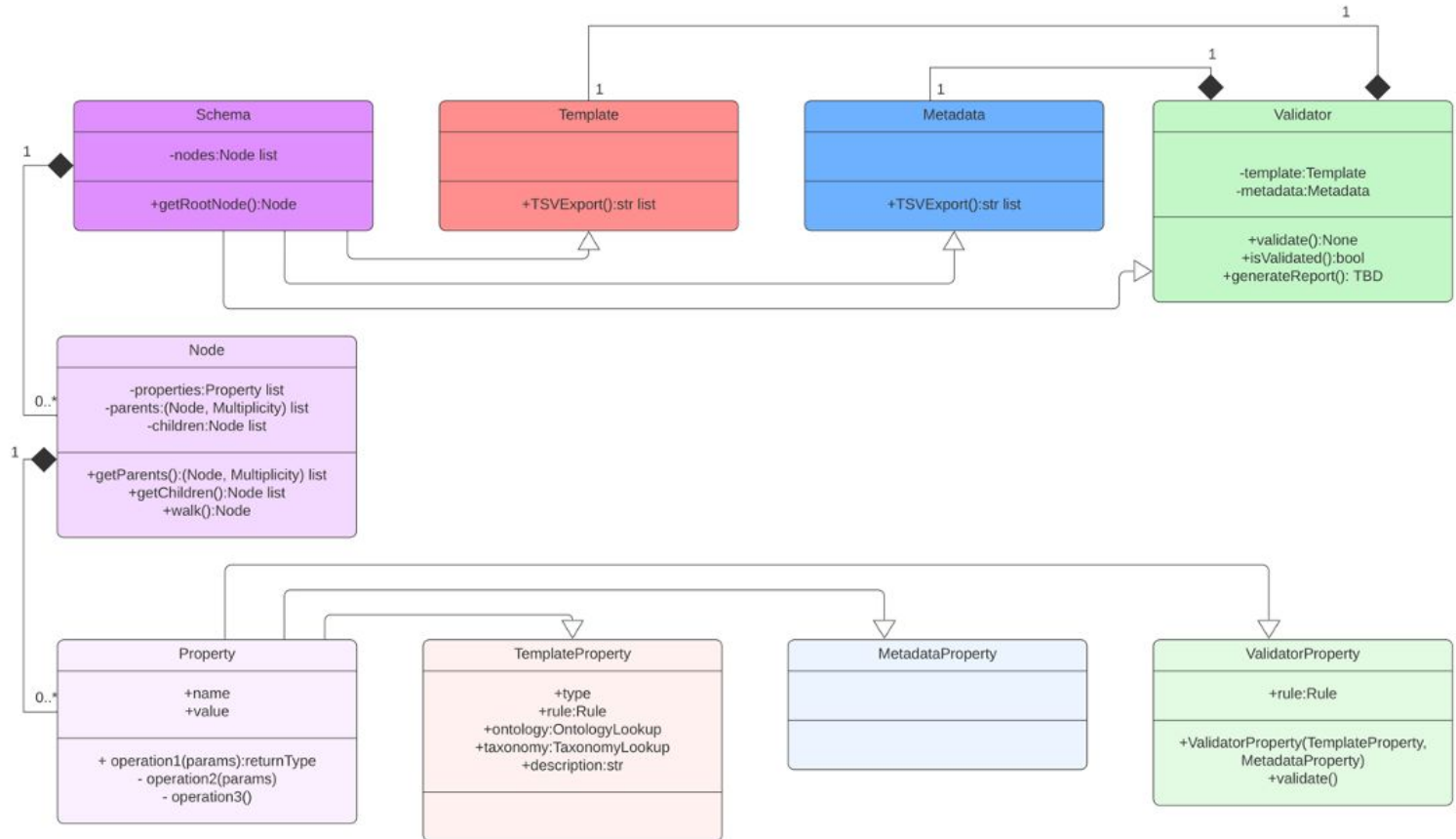
```
1 id: _terms
2
3 28s_16s_ribosomal_rna_ratio:
4   description: >
5     The 28S/18S ribosomal RNA band ratio used to assess the quality of total RNA.
6   termDef:
7     term: "28s/18s Ribosomal RNA Ratio"
8     source: null
9     cde_id: null
10    cde_version: null
11    term_url: null
12
13 a260_a280_ratio:
14   description: >
15     Numeric value that represents the sample ratio of nucleic acid absorbance at 260
16     used to determine a measure of DNA purity.
17   termDef:
18     term: Nucleic Acid Absorbance at 260 And Absorbance at 280 DNA Purity Ratio Valu
19     source: caDSR
20     cde_id: 5432595
21     cde_version: 1.0
22     term_url: "https://cdebrowser.nci.nih.gov/cdebrowserClient/cdeBrowser.html#/sear
23
24 aa_change:
25   description: >
26     Alphanumeric value used to describe the amino acid change for a specific genetic
27     Example: R116Q.
28   termDef:
29     term: Molecular Laboratory Procedure Amino Acid Change Text
30     source: caDSR
```


1. Represent metadata as a **graph**

2. Represent dictionary as a **graph**

3. Combine into **graph** and perform rule application on all properties of each node





- Iterating over all nodes

```
23     def walk(self, revisitNodes=False):
24         visitedNodes = set()
25
26         def bfs(node):
27             if node.getChildren():
28                 for child in node.getChildren():
29                     if not revisitNodes:
30                         if child.name in visitedNodes:
31                             continue
32                         visitedNodes.add(child.name)
33                         yield child
34                         yield from bfs(child)
35             return
36
37         for node in bfs(self._root):
38             yield node
```

- Validate each node and its properties

```
148     def validate(self):
149         # walk nodes
150         # for each node, call validate
151         for node in self.walk():
152             isValid, reasons = node.validate()
153             if not isValid:
154                 self.report(node._input_name, reasons)
155
```

- Parse a dictionary
 - (Excuse the mess)
 - Could be further simplified

```
def parse(self):
    root = self._extractRoot()
    self._schema.setRoot(root)
    current_depth = [root]
    next_depth = []

    while current_depth != []:
        for current_node in current_depth:
            self._schema.nodes[current_node.name] = current_node
            for potential_child in list(self._gen3Dictionary):
                pchild_node = node.Gen3(self._gen3Dictionary[potential_child], self._gen3Dictionary
                [potential_child]["id"])
                logger.debug(f"_____checking node: [_____{pchild_node.name}_____] with potential parent:
                {current_node.name}")
                if pchild_node.isChildOf(current_node):
                    logger.debug(f"found child of {current_node.name}: {pchild_node.name}")

                    logger.debug(f"______{pchild_node.name}_____")
                    pchild_node.parse_properties(self._gen3Dictionary[potential_child]["properties"],
                    self._gen3Dictionary[potential_child]["required"], self._schema._terms, self.
                    _schema._definitions, self._schema._settings)
                    #print(f"found child of {current_node.name}: {pchild_node.name}")
                    current_node.addChild(pchild_node)
                    pchild_node.addParent(current_node)
                    next_depth.append(pchild_node)
                    self._gen3Dictionary.pop(potential_child)
            else:
                current_depth = next_depth
                next_depth = []

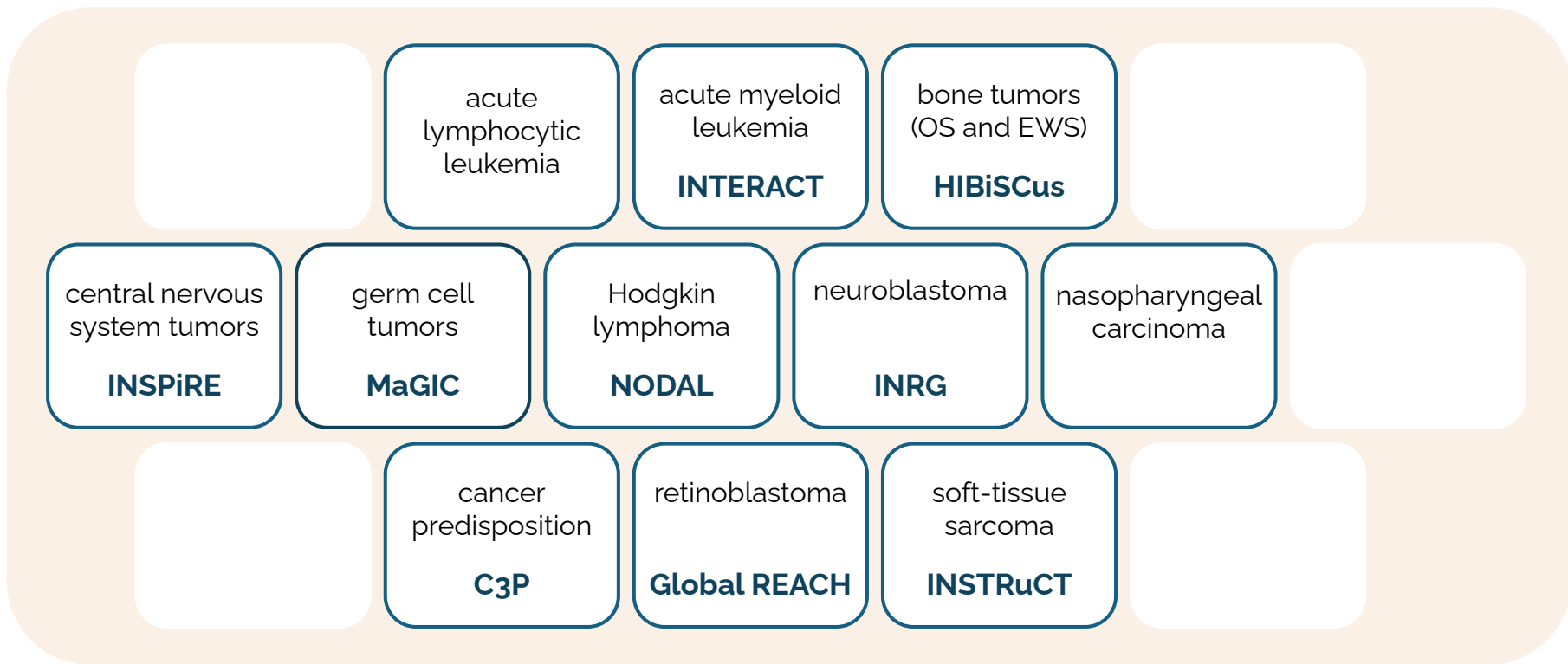
    # do some post processing;
```

Thanks!

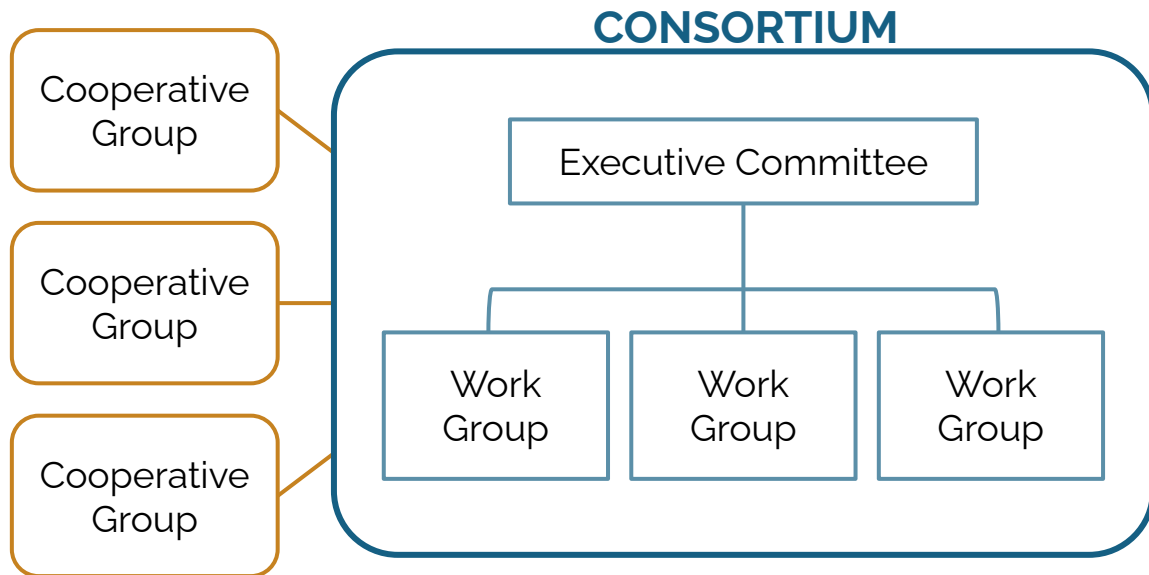
Versioning, migrations, and data release processes in the Pediatric Cancer Data Commons

Brian Furner

The PCDC: a consortium of consortia



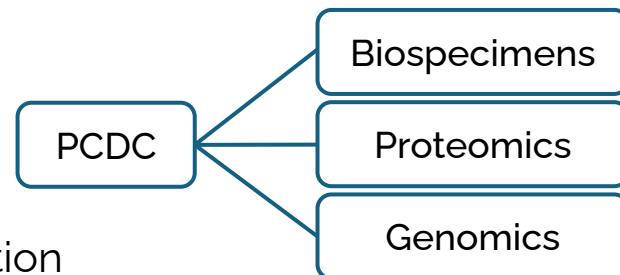
A consortium for each disease group



- drives the science
- creates data dictionary
- harmonizes data
- fuels research and discovery

The PCDC is a clinical data commons

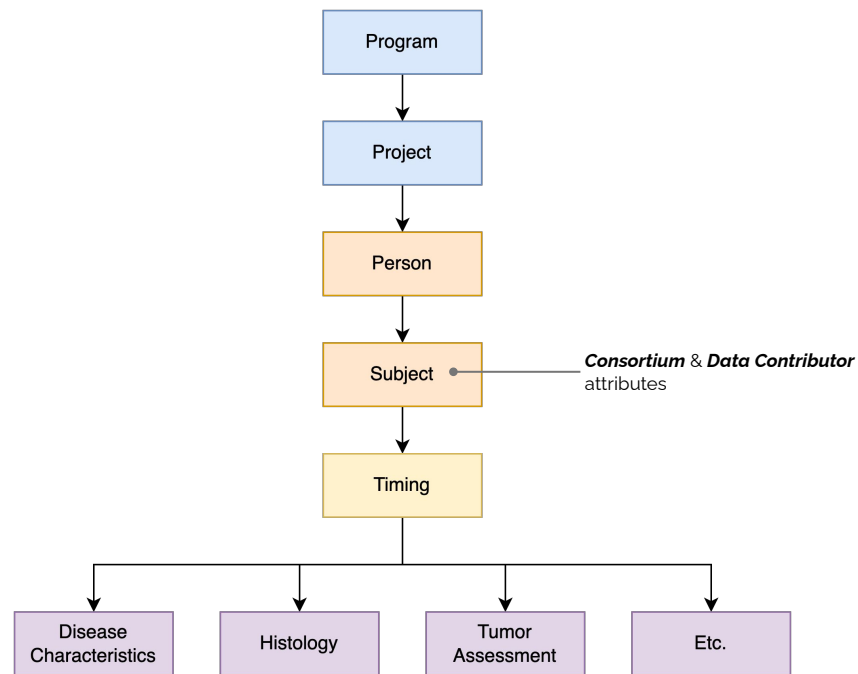
- Clinical data represented include:
 - Demographics
 - Lab values
 - Tumor information
 - Genetic test results
 - Treatment information / Clinical trial information



- Data are sourced from **completed trials, registries, and the EHR**
- **Links** to other data are preserved wherever possible
- A **single, aggregated data model** underpins the whole PCDC

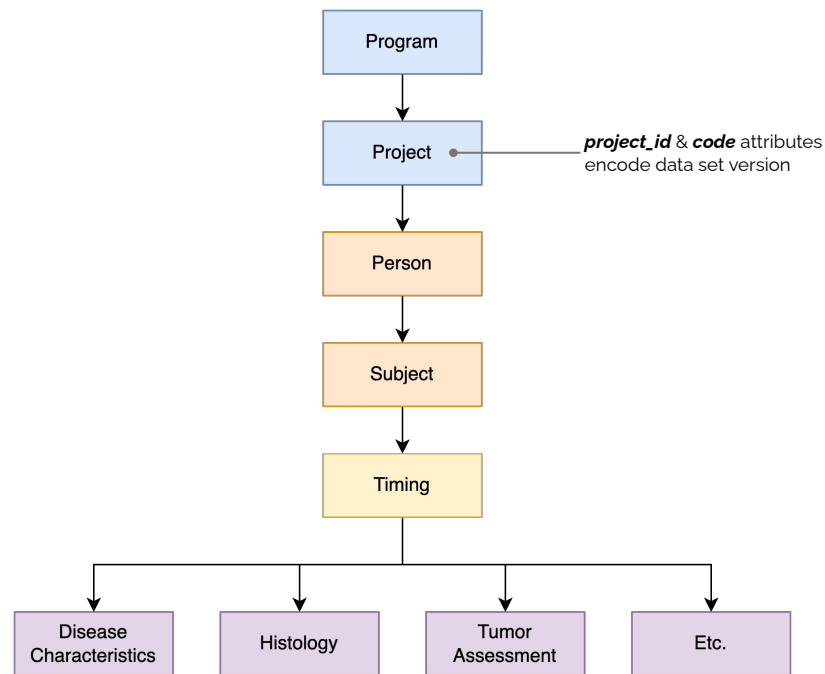
Simplified high-level PCDC model

- Current model has 45 nodes and > 600 properties
- **Person** models a unique individual who may be a **Subject** in one or more research studies, in one or more consortium, and from one or more data contributor
- **Subject** has attributes that hold the associations to a specific consortium and data contributor
- Observations (e.g., **Histology**, **Tumor Assessment**) about a **Subject** are organized in downstream nodes that are related to the **Subject** through an optional **Timing** node



Simplified high-level PCDC model

- Releases occur about once per quarter and include a combination of new / updated records, dictionary changes
- The **entire data set is versioned** at each release
- Data set versions / releases are handled by creating new **Project** records and **all data further down in the graph are (re)loaded** and associated with the new **Project** record
- As a result, new PCDC releases can be time consuming as **records need to be (re)submitted to the graph**
 - Full load of the graph takes **~1.5 days**
 - Any corrections that need to be made during a load can be costly from a timing perspective



Versioning and Migration Process

- We would like to be able to keep 'point-in-time' archival snapshots of the graph
 - Useful for **troubleshooting data change** over time
 - Allows for **reproducibility of analytic data subsets** given to PCDC users
 - While changes between PCDC data set versions are incremental, given our modeling choice, we need to perform **full loads on each release**
- Currently exploring using PFB to support these processes
 - **Export entire graph for archival purposes** rather than multiple concurrent versions in the graph
 - **Import entire graph (or subsets) to 'seed' migrations** rather than submitting all records through the API

Open Discussion

Topic Ideas for Gen3 Community Events

- **Speakers**

- Robert Grossman - Center for Translational Data Science, University of Chicago
- Michael Fitzsimons - Center for Translational Data Science, University of Chicago
- Marion Shadbolt - Australian BioCommons
- Eirian Perkins - New Zealand eScience Infrastructure (NeSI)
- Chris Meyer - Center for Translational Data Science, University of Chicago
- Brian Furner - Data for the Common Good, University of Chicago

- **Gen3 Forum Steering Committee**

- Robert Grossman - Center for Translational Data Science, University of Chicago
- Steven Manos - Australian BioCommons
- Claire Rye - New Zealand eScience Infrastructure
- Plamen Martinov - Open Commons Consortium
- Michael Fitzsimons - Center for Translational Data Science, University of Chicago