GEN3 DATA COMMONS

Data Submission - Perspectives and solutions from different Gen3 systems

Gen3 Community Forum
10 July 2024

THE UNIVERSITY OF CHICAGO
CENTER FOR TRANSLATIONAL DATA SCIENCE

OHSU

Australian BioCommons

NeSI
New Zealand eScience Infrastructure

The Agenda

- Introduction
- Data Submission into the Aotearoa Genomic Data Repository - Nathalie Giraudon, Claire Rye, New Zealand eScience Infrastructure (NeSI)
- COVID-19 Data: 7 Million Records in PRC - Plamen Martinov, Open Commons Consortium
- MIDRC: Ingesting Medical Images from Multiple Contributors - Chris Meyer, Center for Translational Data Science, University of Chicago
- g3t: Gen3 Tracker – User Driven Submissions - Jordan Lee and Liam Beckman, Oregon Health and Science University
- Data Modelling in Gen3 - Joshua Harris, Australian BioCommons

Data submission into the Aotearoa Genomic Data Repository

Nathalie Giraudon, Claire Rye
New Zealand eScience Infrastructure (NeSI)

Data submission into a Gen3

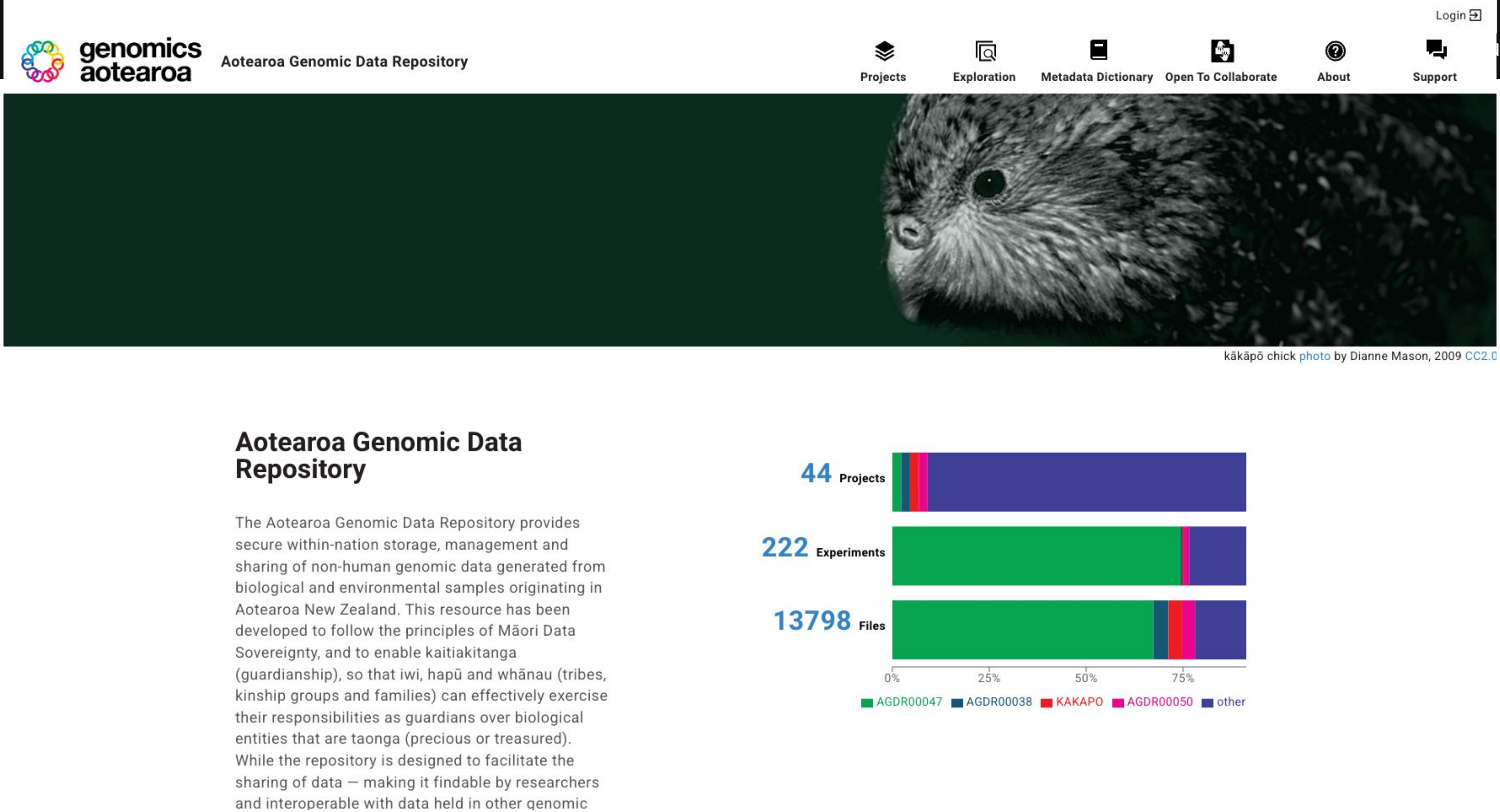
A Case study of the Aotearoa Genomic Data Repository

Nathalie Giraudon, Claire Rye
NeSI / University of Auckland

10/11 July 2024

Agenda

- What is AGDR (Aotearoa Genomic Data Repository)?
- Process of data submission
  - metadata spreadsheet
  - metadata validator
    - Reasoning
    - Principles
  - Ingestion
- Demonstration
- Issues
- Next steps



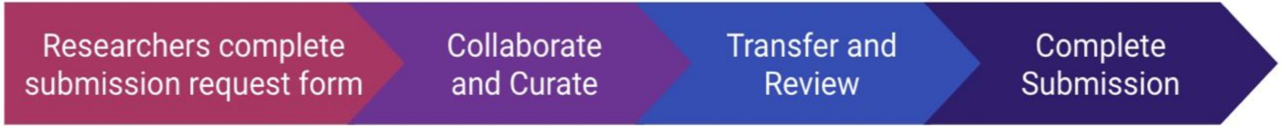
kākāpō chick photo by Dianne Mason, 2009 CC2.0

Aotearoa Genomic Data Repository

The Aotearoa Genomic Data Repository provides secure within-nation storage, management and sharing of non-human genomic data generated from biological and environmental samples originating in Aotearoa New Zealand. This resource has been developed to follow the principles of Māori Data Sovereignty, and to enable kaitiakitanga (guardianship), so that iwi, hapū and whānau (tribes, kinship groups and families) can effectively exercise their responsibilities as guardians over biological entities that are taonga (precious or treasured). While the repository is designed to facilitate the sharing of data — making it findable by researchers and interoperable with data held in other genomic

https://data.agdr.org.nz/

Data Submission



Researchers complete submission request form

Collaborate and Curate

Transfer and Review

Complete Submission

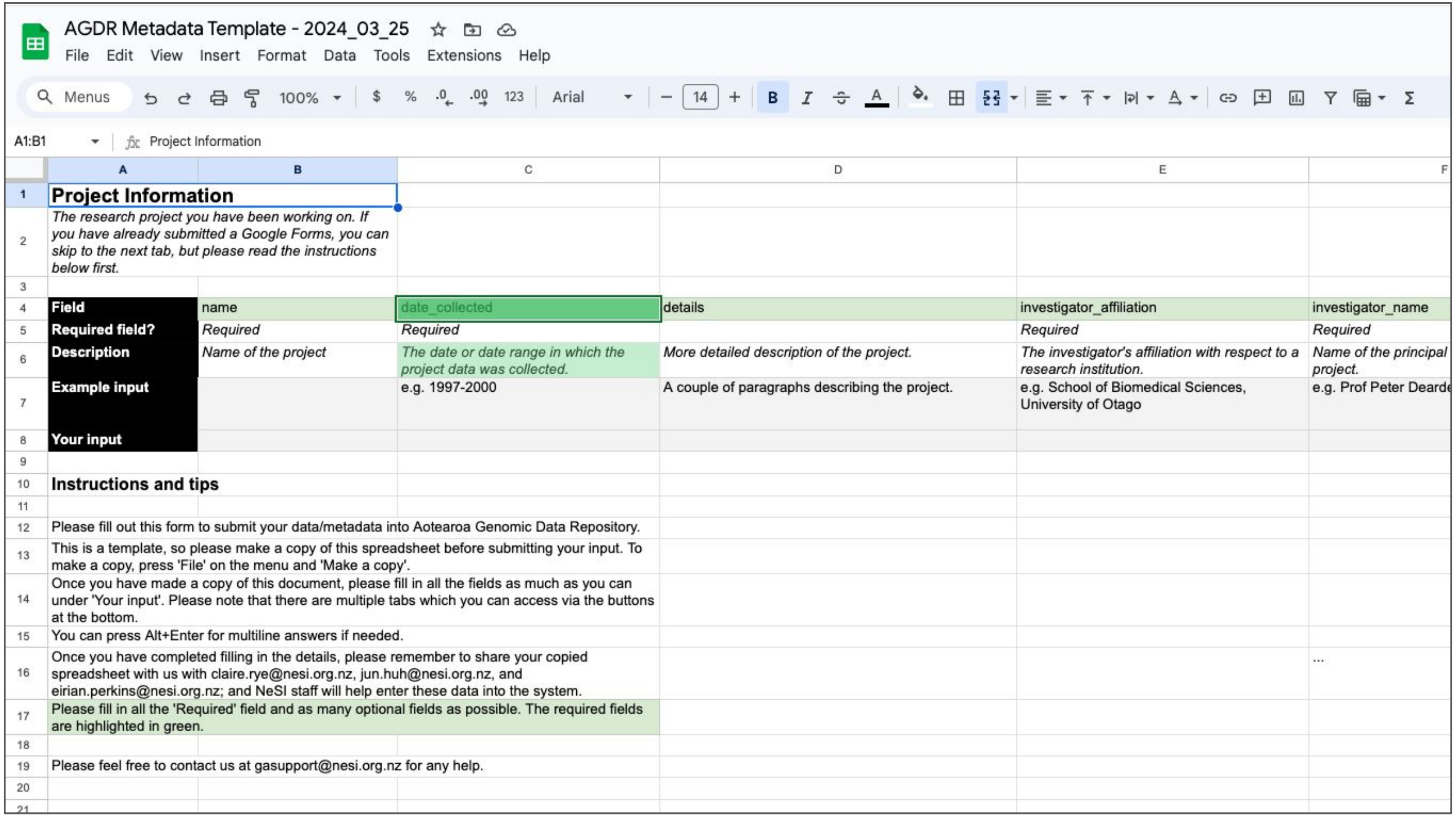
- Suitability assessment

- Metadata template spreadsheet
- Metadata standard, community built, MIxS compliant

- Upload data
- Metadata reviewed

- Project viewable in AGDR
- DOI provided for use in publications

Spreadsheet template



Project Information

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.

Field	name	date_collected	details	investigator_affiliation	investigator_name
Required field?	*Required*	*Required*		*Required*	*Required*
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.*	*Name of the principal project.*
Example input		e.g. 1997-2000	A couple of paragraphs describing the project.	e.g. School of Biomedical Sciences, University of Otago	e.g. Prof Peter Dearde
Your input					

Instructions and tips

Please fill out this form to submit your data/metadata into Aotearoa Genomic Data Repository.

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'.

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.

You can press Alt+Enter for multiline answers if needed.

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.

...

Please fill in all the 'Required' field and as many optional fields as possible. The required fields are highlighted in green.

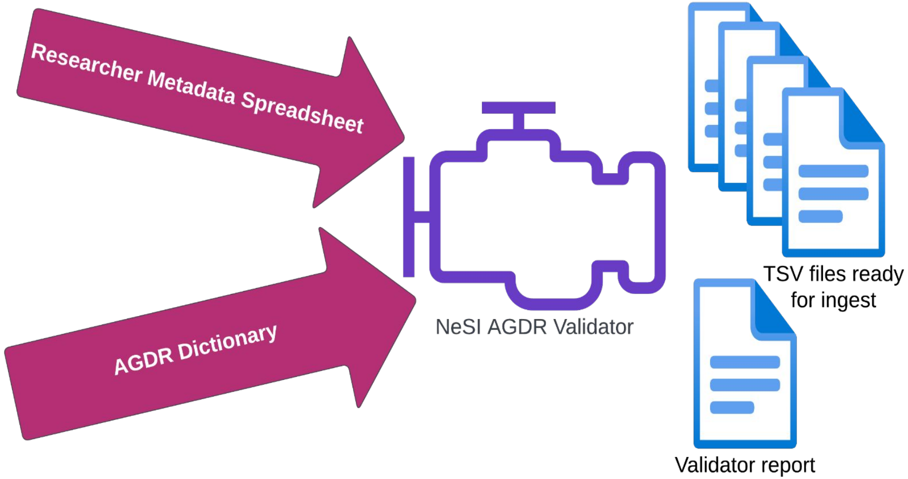
Please feel free to contact us at gasupport@nesi.org.nz for any help.

Metadata Validation

Reasons:

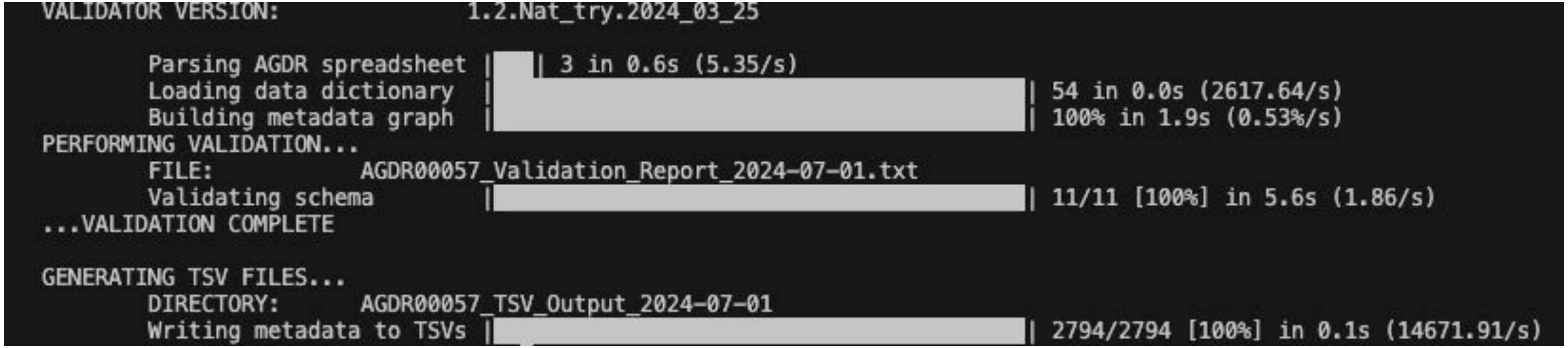
- Simplification (no need to know the dictionary, submitter_id meaning not obvious…)
- Data files are not loaded with the Gen3 client tool
- Consistency checks for the values
- Early checks before ingest

Principles:



Demonstration of the validator/ingest

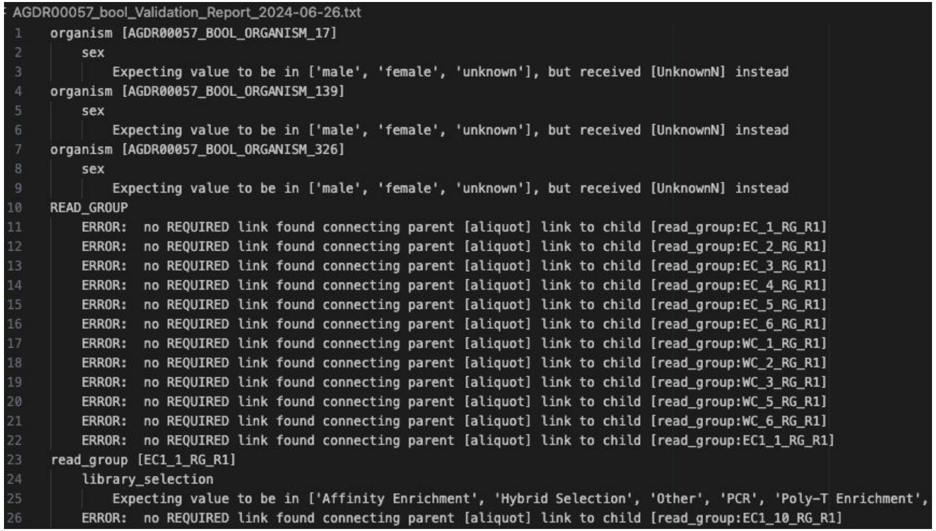
- Small dataset with spreadsheet and dictionary errors
- Small dataset with no validator errors
  - Then ingest of the tsv files on our test system
- Large dataset and ingest



```
VALIDATOR VERSION:              1.2.Nat_try.2024_03_25

        Parsing AGDR spreadsheet |███| 3 in 0.6s (5.35/s)
        Loading data dictionary  |████████████████████████████████████████| 54 in 0.0s (2617.64/s)
        Building metadata graph  |████████████████████████████████████████| 100% in 1.9s (0.53%/s)
PERFORMING VALIDATION...
        FILE:           AGDR00057_Validation_Report_2024-07-01.txt
        Validating schema        |████████████████████████████████████████| 11/11 [100%] in 5.6s (1.86/s)
...VALIDATION COMPLETE

GENERATING TSV FILES...
        DIRECTORY:      AGDR00057_TSV_Output_2024-07-01
        Writing metadata to TSVs |████████████████████████████████████████| 2794/2794 [100%] in 0.1s (14671.91/s)
```

Issues

- Ingest of project via UI
- Problem with large datasets error
- Special characters support - issues (Excel?)
- Robustness of the validator
  - true/TRUE/’true -> can only ingest boolean in lower case…

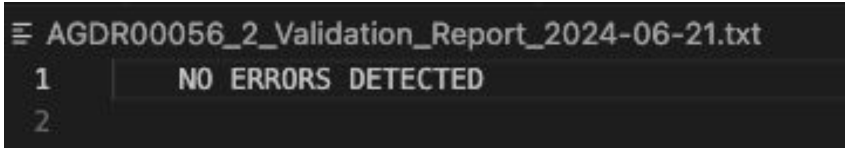


AGDR00057_bool_Validation_Report_2024-06-26.txt

```
organism [AGDR00057_BOOL_ORGANISM_17]
    sex
        Expecting value to be in ['male', 'female', 'unknown'], but received [UnknownN] instead
organism [AGDR00057_BOOL_ORGANISM_139]
    sex
        Expecting value to be in ['male', 'female', 'unknown'], but received [UnknownN] instead
organism [AGDR00057_BOOL_ORGANISM_326]
    sex
        Expecting value to be in ['male', 'female', 'unknown'], but received [UnknownN] instead
READ_GROUP
    ERROR:  no REQUIRED link found connecting parent [aliquot] link to child [read_group:EC_1_RG_R1]
    ERROR:  no REQUIRED link found connecting parent [aliquot] link to child [read_group:EC_2_RG_R1]
    ERROR:  no REQUIRED link found connecting parent [aliquot] link to child [read_group:EC_3_RG_R1]
    ERROR:  no REQUIRED link found connecting parent [aliquot] link to child [read_group:EC_4_RG_R1]
    ERROR:  no REQUIRED link found connecting parent [aliquot] link to child [read_group:EC_5_RG_R1]
    ERROR:  no REQUIRED link found connecting parent [aliquot] link to child [read_group:EC_6_RG_R1]
    ERROR:  no REQUIRED link found connecting parent [aliquot] link to child [read_group:WC_1_RG_R1]
    ERROR:  no REQUIRED link found connecting parent [aliquot] link to child [read_group:WC_2_RG_R1]
    ERROR:  no REQUIRED link found connecting parent [aliquot] link to child [read_group:WC_3_RG_R1]
    ERROR:  no REQUIRED link found connecting parent [aliquot] link to child [read_group:WC_5_RG_R1]
    ERROR:  no REQUIRED link found connecting parent [aliquot] link to child [read_group:WC_6_RG_R1]
    ERROR:  no REQUIRED link found connecting parent [aliquot] link to child [read_group:EC1_1_RG_R1]
read_group [EC1_1_RG_R1]
    library_selection
        Expecting value to be in ['Affinity Enrichment', 'Hybrid Selection', 'Other', 'PCR', 'Poly-T Enrichment',
    ERROR:  no REQUIRED link found connecting parent [aliquot] link to child [read_group:EC1_10_RG_R1]
```

Next steps

- Update of Elastic Search/Etlmapping versions
- Last validator improvements before the first release
    - Support of multi links
- Validator release and training with the researchers
- Automation of the ingest via API



AGDR00056_2_Validation_Report_2024-06-21.txt

```
NO ERRORS DETECTED
```

Thank you!

- GA Team:
  - Bioinformatics project: Libby Liggins (Massey), Rudi Brauning (AgResearch), Mik Black (Otago), Tracey Godfery (Otago), Tanis Goodwin (Otago)
- NeSI team (Auckland):
  - Jun Huh, Eirian Perkins, Claire Rye, Nathalie Giraudon, Rui Chen (Carvin)
- Former team members:
  - Miles Benton (ESR/Oxford Nanopore), Ben Te Aika (Otago), Ben Curran (Auckland), Brian Flaherty, Thomas Berger, Kenny Zhao (NeSI).

COVID-19 Data: Case Study of submitting 7 Million Subjects Health Data in PRC

Plamen Martinov
Mikisha Patel
Urvi Sheth
Open Commons Consortium

Plamen Martinov,
Chief Technology and Information Security Officer

Background

A Collaborative Data Ecosystem to Improve Outcomes for COVID-19

a. COVID19 shut the world down

b. OCC expertise in Data Commons and Data Meshes took charge

c. Working shoulder to shoulder with the University of Chicago, Center for Transportation Data Science we created, first of it's kind Chicagoland region COVID-19 Data Commons

d. Using the expertise from all teams we forged agreements (aka Common Legal Agreements) with regional health care organizations to bring valuable data for ongoing research

e. Using the FAIR model as a framework we created a secure space using Gen3 Data Commons for organizing and sharing data related to COVID-19

Background

Progress to date:

a. We started and defined a data dictionary that help answers questions related to the Case Fatality Ratio eventually making this data public

b. We integrated a system developed by SIU that tracked mobility of COVID-19 cases

c. We are now working on Long COVID-19 analysis systemantics through a devolved governance structure

d. Now we have 7 million subject level records from members of the group to continue the ongoing research

Mikisha Patel,
Technical Project Manager

Data Submission Process

How OCC works with different organizations & Universities to collect data

- Complete Contracts & Legal Documentation
- IRB Approval
- Establishing a clear Data Dictionary
- Onboarding member to PRC
- Upload/submit data to PRC Data Commons

Complete Contracts & Legal Documentation

Establish contracts and legal documentation with member organization and universities for the data submission https://pandemicresponsecommons.org/governance/legal-agreements/

Currently we have 4 member organization who submits the data quarterly:

- Rush
- Northshore
- UIC
- UChicago

IRB Approval

Each organization and university must obtain IRB approval to upload the data fields

Align to the Data Dictionary

- Coordinate with the different organizations and universities to align on a data dictionary to be collected
- Chicagoland COVID-19 Commons dictionary has 41 nodes and 1245 properties
- The data Dictionary can be viewed here: https://chicagoland.pandemicresponsecommons.org/DD

Member Onboarding to PRC

- Providing access to PRC Data Commons
- OCC will create a project to host the data, and provide access to the the members accordingly

Urvi Sheth,

Sr. Software Engineer

Upload/Submit Data to PRC Data Commons

- Secure Data Submission or Retrieval
  - Ensuring secure methods for data submission or retrieval.
- Data Pre-processing and Validation
  - Cleaning and validating data to meet data commons standards.
- Data Modeling
  - Creating structured models based on data dictionary.
- Stakeholder Training (optional)
  - Offering training on data upload.
- Data Upload
  - Uploading data to designated nodes within the project.
- Data Utilization and Visualization
  - Enabling data access for analysis and various visualizations in Gen3 Commons.

MIDRC Data Ingestion

Wrangling Imaging Data from Multiple Sources

Chris Meyer, PhD
Center for Translational Data Science
The University of Chicago

Gen3 Community Forum
10 July 2024

THE UNIVERSITY OF CHICAGO
CENTER FOR TRANSLATIONAL DATA SCIENCE

MIDRC
MEDICAL IMAGING AND DATA RESOURCE CENTER.

MIDRC Initiative Overview

MIDRC was launched in 2020 and aims to accelerate medical machine learning innovation by providing a high-quality, curated data resource, which includes medical imaging studies and associated clinical data.

MIDRC is funded by NIBIB, is hosted at UChicago, and is co-led by ACR, RSNA, and AAPM.

Most data to date are COVID-19-related clinical and imaging studies, but MIDRC is expanding to other diseases, like cancer and long COVID.

https://midrc.org

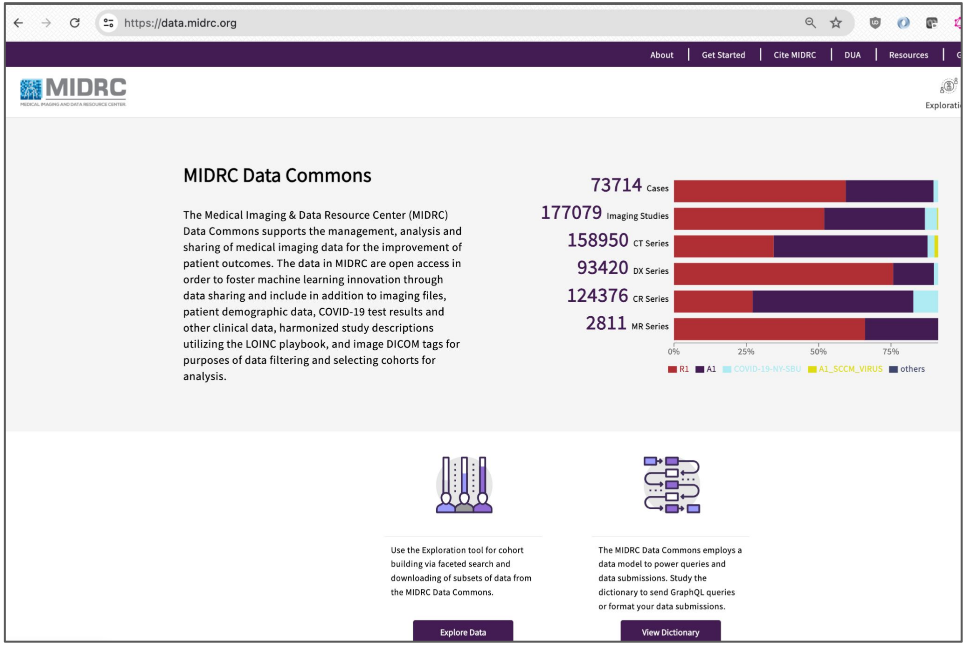
The MIDRC Data Commons

MIDRC operates four Gen3 environments:

- Production
  - **data.midrc.org** (Open)
  - **validate.midrc.org** (Sequestered)
- Staging
  - **staging.midrc.org** (Open Staging)
  - **validatestaging.midrc.org** (Sequestered Staging)

The **open data are for training** AI algorithms and the **sequestered data are for testing** against a demographically balanced subset.

Data are ingested in staging environments then copied to production after QC.



MIDRC Data Commons

The Medical Imaging & Data Resource Center (MIDRC) Data Commons supports the management, analysis and sharing of medical imaging data for the improvement of patient outcomes. The data in MIDRC are open access in order to foster machine learning innovation through data sharing and include in addition to imaging files, patient demographic data, COVID-19 test results and other clinical data, harmonized study descriptions utilizing the LOINC playbook, and image DICOM tags for purposes of data filtering and selecting cohorts for analysis.

Use the Exploration tool for cohort building via faceted search and downloading of subsets of data from the MIDRC Data Commons.

The MIDRC Data Commons employs a data model to power queries and data submissions. Study the dictionary to send GraphQL queries or format your data submissions.

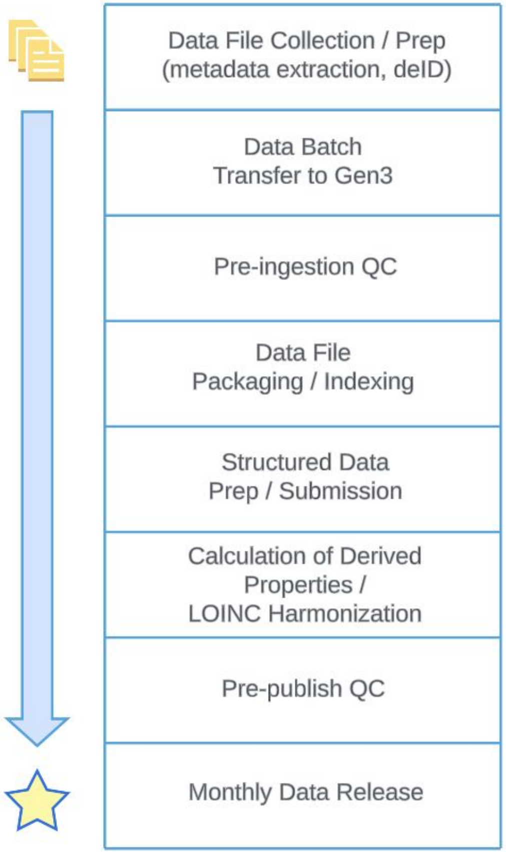
https://data.midrc.org

MIDRC Data Ingestion Overview

MIDRC publishes new batches of data on a **monthly release cadence** (data, services, and config changes).

The Gen3 team QC's data on initial receipt and before it's published (copied from staging environment to production).

SOP documents exist for all of these processes, and where we can, **processes are scripted**, e.g., in Python or Jupyter Notebooks (Data QC, preparation/submission, and release).



Data File Collection / Prep (metadata extraction, deID)

Data Batch Transfer to Gen3

Pre-ingestion QC

Data File Packaging / Indexing

Structured Data Prep / Submission

Calculation of Derived Properties / LOINC Harmonization

Pre-publish QC

Monthly Data Release

Collection of Clinical and Imaging Data

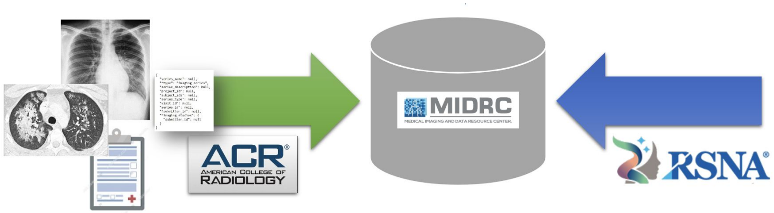
GEN3 DATA COMMONS

Data Continuously Flows into **2 Primary Data Intake Portals** from Contributing Medical Sites:

- American College of Radiology (ACR) **C**OVID **I**maging **R**esearch **R**egistry (**CIRR**)
- **R**adiological Society of North America (RSNA) **I**nternational **C**OVID-19 **O**pen **R**adiology **D**atabase (**RICORD**)

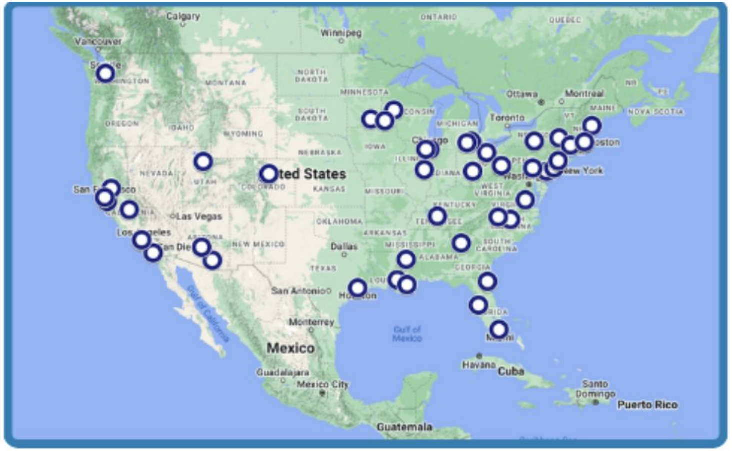
The RSNA and ACR teams:

- Collect clinical and imaging data from medical centers
- De-Identify structured EHR data and Images
- Provide Gen3 team access to batches of data for ingestion into the MIDRC data commons.



MIDRC
MEDICAL IMAGING AND DATA RESOURCE CENTER
ACR AMERICAN COLLEGE OF RADIOLOGY
RSNA

Multiple Pathways for Contribution



Contributions coming from 23 states

https://www.midrc.org/donate

Data Modeling and Harmonization

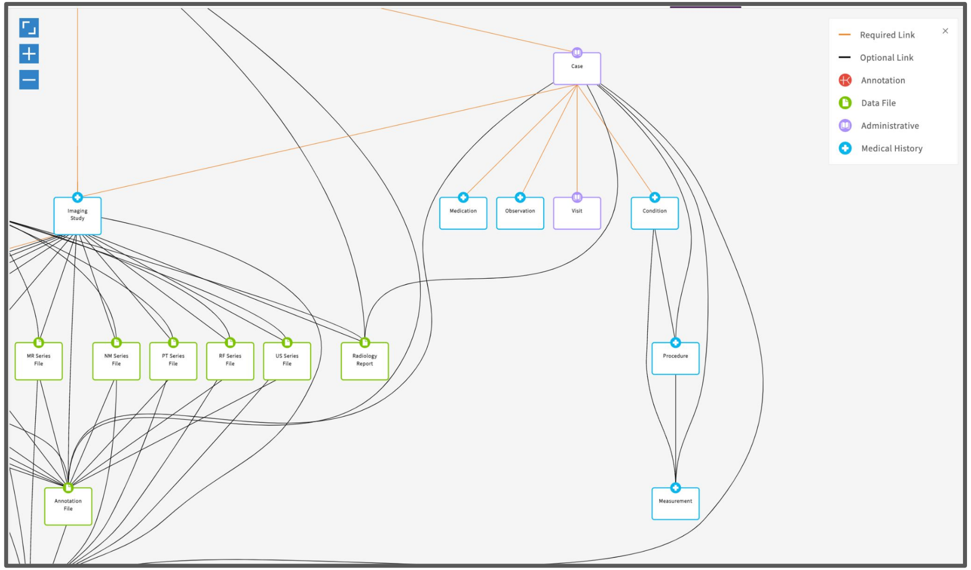
MIDRC has a subcommittee that collaboratively develops the Gen3 graph data model: **D**ata **S**tandards and **I**nformation **T**echnology (**DSIT**).

ACR and RSNA are members and work closely with the Gen3 team to **implement a data model that best supports queries for cohort building** using patient EHR and image DICOM metadata.

ACR and RSNA extract the DICOM metadata from batches of images and organize it into Gen3 submission TSVs that conform to the data model.

Associated clinical data is similarly extracted from EHR platforms and organized into clinical TSVs.

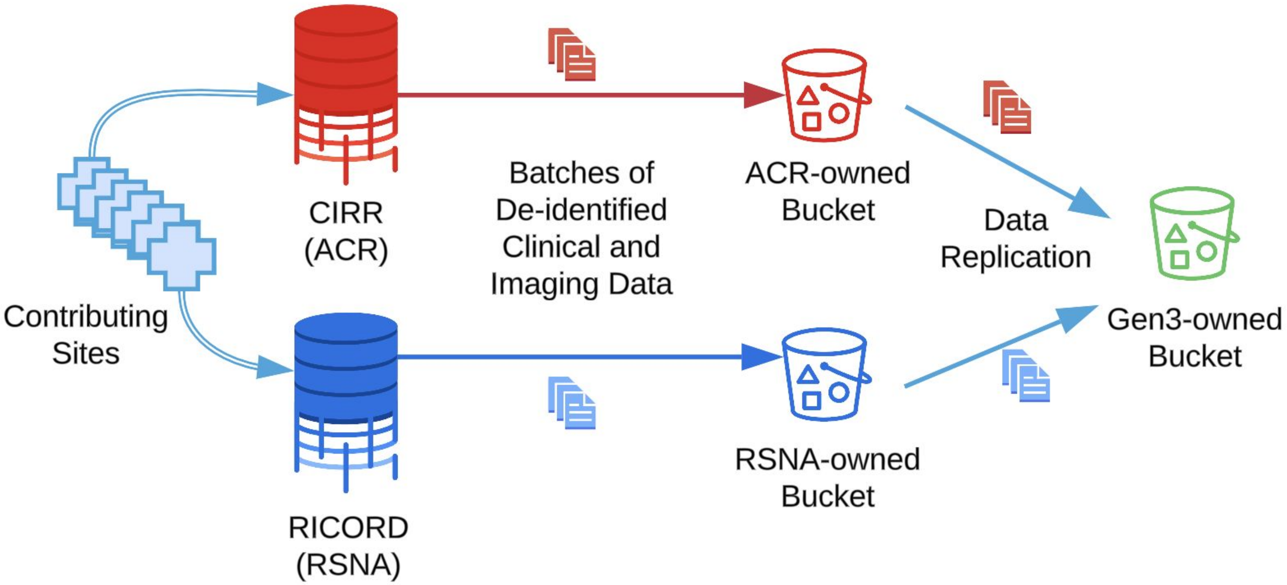
https://www.midrc.org/subcommittees



Data Dictionary Viewer:
https://data.midrc.org/dd

Dictionary in GitHub:
https://github.com/uc-cdis/midrc_dictionary

Data Transfer to Gen3



- RSNA and ACR periodically make batches of de-identified data available to the Gen3 Team.
- A batch consists of: structured data/submission TSVs, image files, and an image manifest.
- The Gen3 team copies each batch in order to ingest it into the MIDRC data commons.

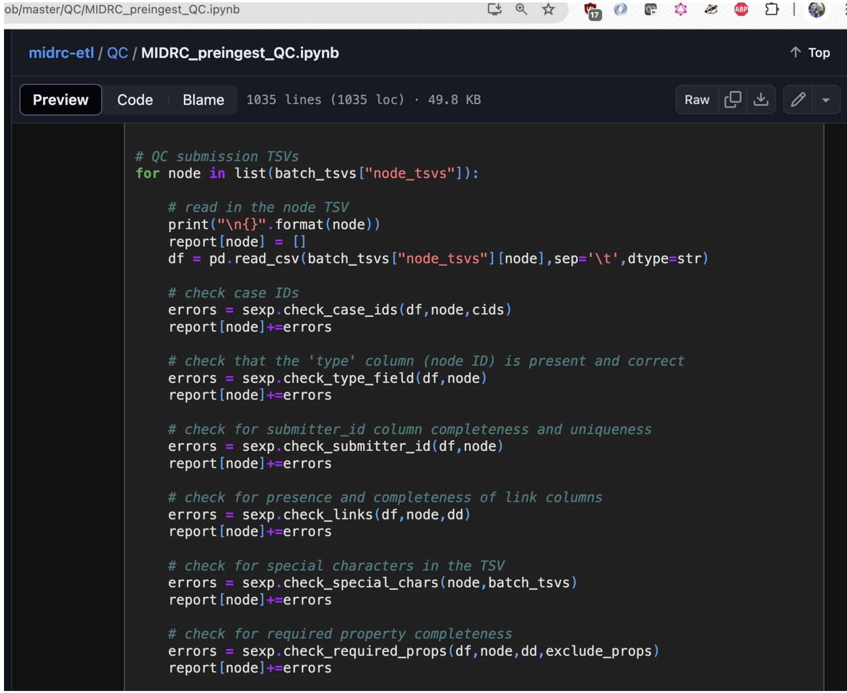
Pre-ingestion QC

Before data are ingested, Gen3 runs a “Pre-ingestion QC Checklist” to ensure completeness and proper formatting.

Checks are implemented in a Jupyter notebook to:

- Confirm reported numbers of patients, imaging studies, and files (in the MIDRC External Gen3 Data Release Tracker) match data received.
- Check that all required data fields are present and complete; report on completeness of optional data fields.
- Check submission TSV formatting.

If the data batch fails any checks, Gen3 notifies the data contributor and requests the batch be corrected.


ob/master/QC/MIDRC_preingest_QC.ipynb
midrc-etl / QC / MIDRC_preingest_QC.ipynb
Top
Preview
Code
Blame
1035 lines (1035 loc) · 49.8 KB
Raw

```python
# QC submission TSVs
for node in list(batch_tsvs["node_tsvs"]):

    # read in the node TSV
    print("\n{}".format(node))
    report[node] = []
    df = pd.read_csv(batch_tsvs["node_tsvs"][node],sep='\t',dtype=str)

    # check case IDs
    errors = sexp.check_case_ids(df,node,cids)
    report[node]+=errors

    # check that the 'type' column (node ID) is present and correct
    errors = sexp.check_type_field(df,node)
    report[node]+=errors

    # check for submitter_id column completeness and uniqueness
    errors = sexp.check_submitter_id(df,node)
    report[node]+=errors

    # check for presence and completeness of link columns
    errors = sexp.check_links(df,node,dd)
    report[node]+=errors

    # check for special characters in the TSV
    errors = sexp.check_special_chars(node,batch_tsvs)
    report[node]+=errors

    # check for required property completeness
    errors = sexp.check_required_props(df,node,dd,exclude_props)
    report[node]+=errors
```

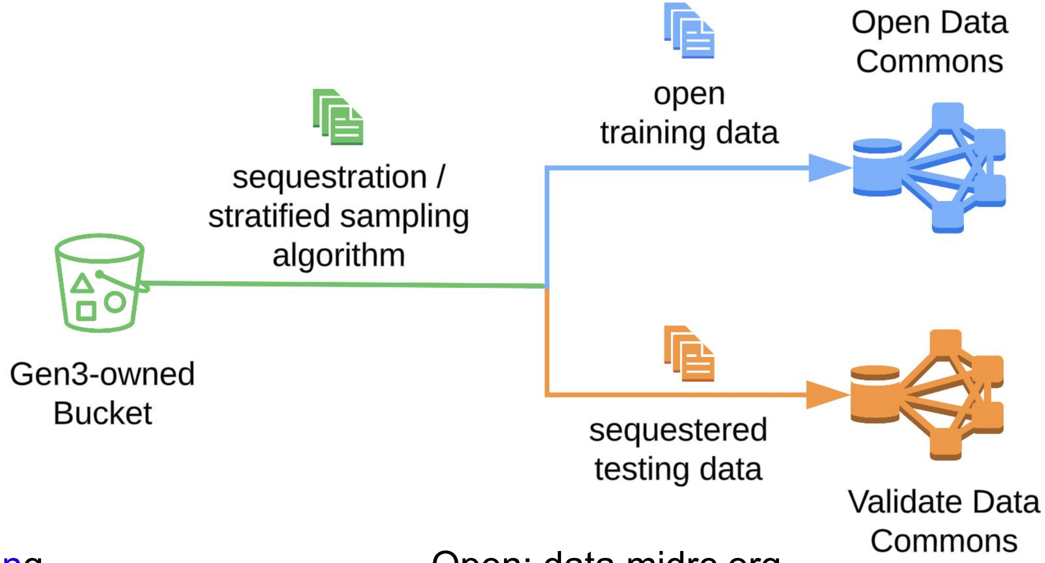
https://github.com/uc-cdis/midrc-etl/blob/master/QC/MIDRC_preingest_QC.ipynb

Patient Sequestration

GEN3 DATA COMMONS

Before ingestion, new patients are split between the **Open (80%)** and **Sequestered (20%)** commons by performing a **stratified sampling algorithm** that attempts to create patient cohorts that are balanced with respect to:

- Patient
  - Age
  - Ethnicity
  - Race
  - Sex
  - Care Site ID
  - COVID-19 status
- Imaging Study
  - Modality
  - Description
  - Body Part Examined



https://github.com/MIDRC/Stratified_Sampling
https://doi.org/10.1117/1.JMI.10.6.064501

Open: data.midrc.org
Sequestered: validate.midrc.org

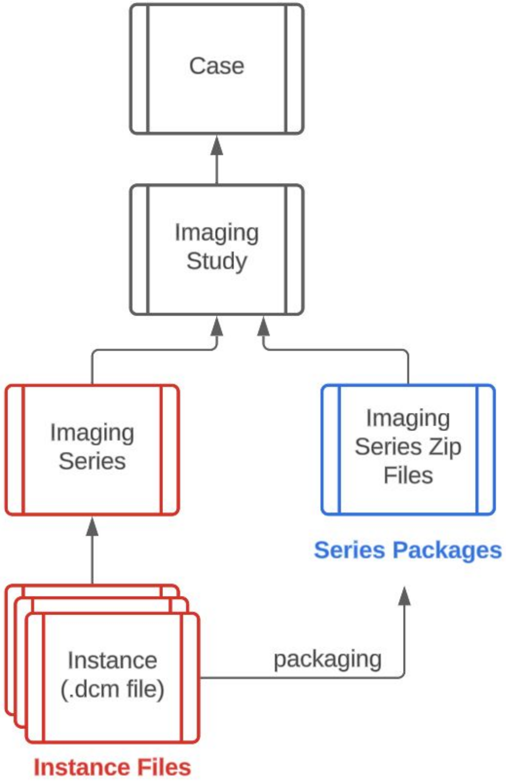
Image File Packaging and Indexing

MIDRC data is in DICOM format, the base-level of which is the image “instance”, which is a .DCM file.

For an x-ray, instances are single images, but for volumetric imaging (MR, CT, etc.) image instances are “slices” in a 3-dimensional image stack comprised of hundreds of instances.

In order to accelerate searches and download speeds for volumetric imaging modalities, **image instances are packaged into series-level zip files** and the zip files are indexed in indexd.

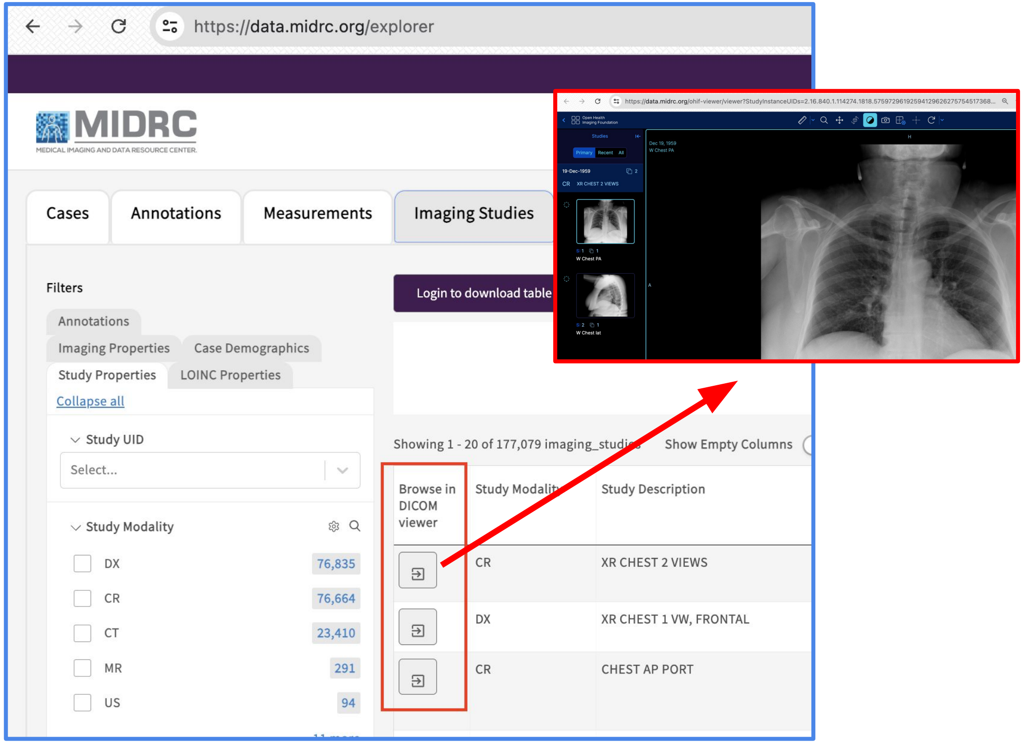
https://github.com/uc-cdis/midrc-etl/tree/master/packaging



Upload Instances to DICOM Viewer Server

Imaging studies in the MIDRC data explorer feature a button that links to a page where imaging series can be viewed in the OHIF DICOM Viewer.

In order for this to work, the image instance files are copied to an Orthanc Server, which organizes the instances (slices) into series and studies for viewing.



https://data.midrc.org/explorer

MIDRC
MEDICAL IMAGING AND DATA RESOURCE CENTER

Filters

Annotations | Imaging Properties | Case Demographics | Study Properties | LOINC Properties

Collapse all

Study UID

Select...

Study Modality

□ DX	76,835
□ CR	76,664
□ CT	23,410
□ MR	291
□ US	94

Showing 1 - 20 of 177,079 imaging_studies

Browse in DICOM viewer	Study Modality	Study Description
	CR	XR CHEST 2 VIEWS
	DX	XR CHEST 1 VW, FRONTAL
	CR	CHEST AP PORT

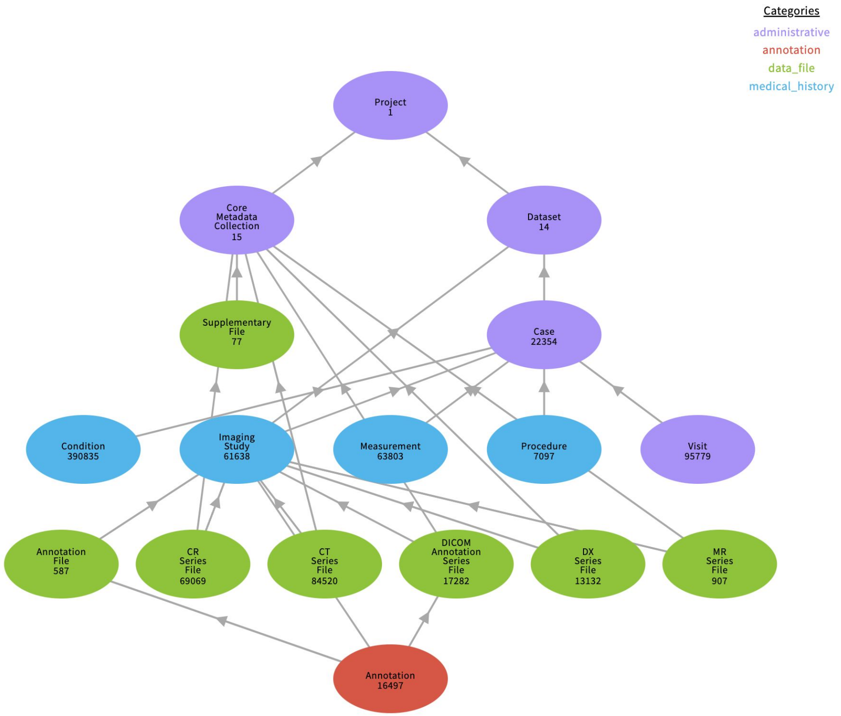
https://data.midrc.org/ohif-viewer/viewer?StudyInstanceUIDs=*<imaging_study.submitter_id>*

Structured Data Submission

Once the image files are packaged and indexed, image package GUIDs are joined to the imaging series TSVs and the structured data TSVs are submitted to the graph via sheepdog using the Gen3SDK “Submission” class function: `Gen3Submission.submit_file().`

- Retries API requests on service failures.
- Returns error messages for troubleshooting
- Returns lists of records by success / failure for faster and simpler retries / resubmissions.

https://github.com/uc-cdis/gen3sdk-python/blob/dbf607b4e91263ea435be27fefedd42fb83daa42/gen3/submission.py#L509



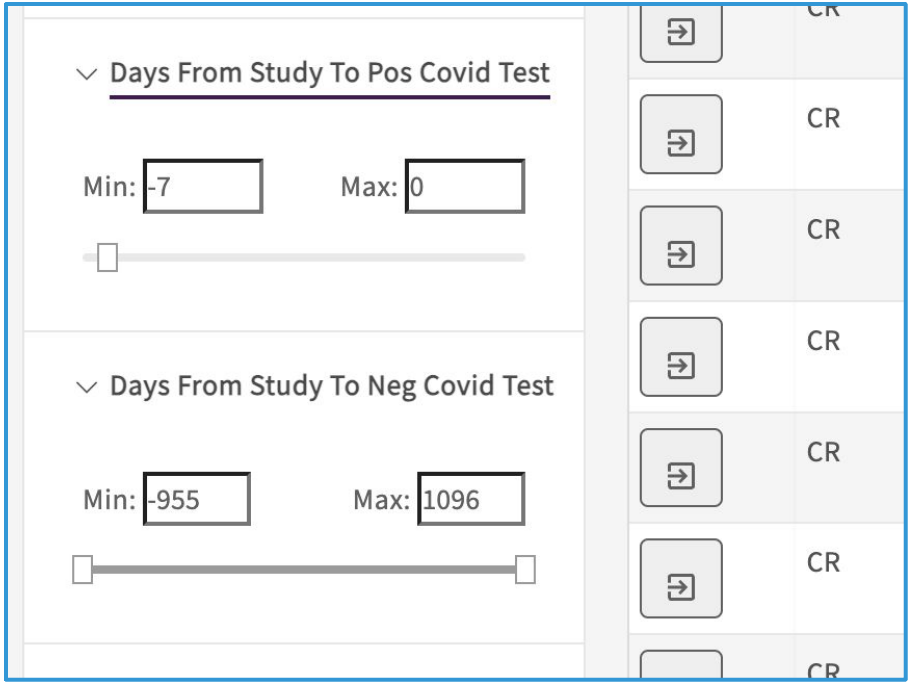
https://data.midrc.org/Open-A1

Calculation of Derived Properties

Certain properties in the MIDRC data model are derived from the raw data and these are calculated and submitted for all relevant records in Staging prior to release via Jupyter Notebook.

Two examples are the number of days between each imaging study and a positive or negative COVID test:

- days_from_study_to_neg_covid_test
- days_from_study_to_pos_covid_test



Days From Study To Pos Covid Test

Min: -7 Max: 0

Days From Study To Neg Covid Test

Min: -955 Max: 1096

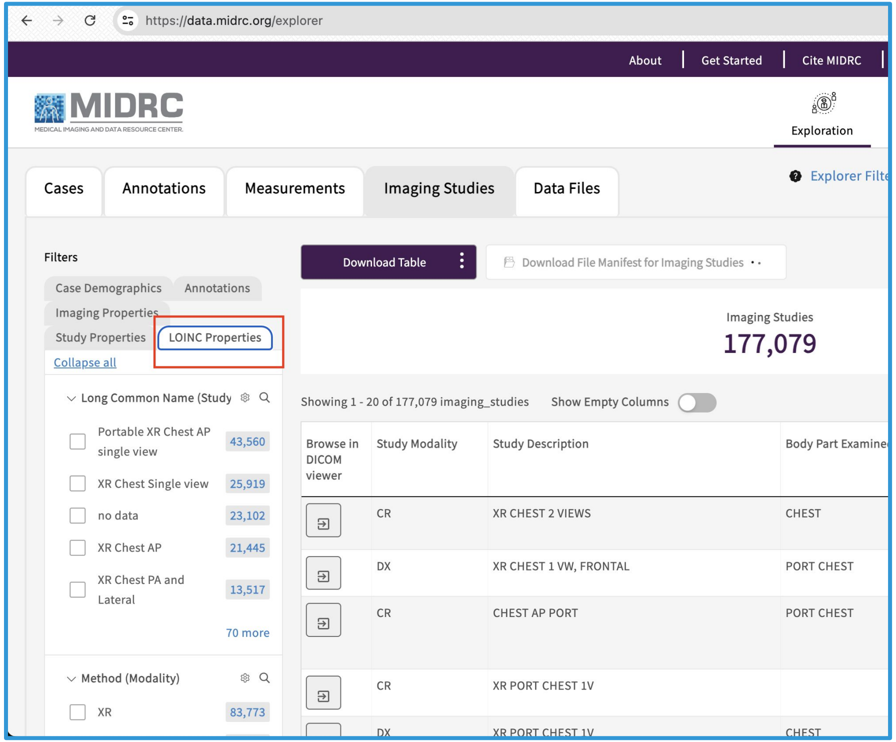
https://github.com/uc-cdis/midrc-etl/blob/master/temporal/calculate_days_from_study_to_covid_test.ipynb
https://data.midrc.orc/explorer

LOINC Mapping / Harmonization

The **D**ata **Q**uality and **H**armonization Subcommittee (**DQH**) has used the LOINC Standard to harmonize over 1,700 disparate imaging study descriptions to only 75 LOINC codes, which encompass the following:

- Study Description
- Modality
- Contrast Indicator
- Body Part Examined

Prior to release of new data to production environments, we perform LOINC mapping and sheepdog update via a Jupyter Notebook.



https://data.midrc.org/explorer

About | Get Started | Cite MIDRC |

MIDRC
MEDICAL IMAGING AND DATA RESOURCE CENTER

Exploration

Cases | Annotations | Measurements | Imaging Studies | Data Files

Explorer Filte

Filters

Case Demographics | Annotations | Imaging Properties | Study Properties | LOINC Properties

Collapse all

Download Table

Download File Manifest for Imaging Studies

Imaging Studies
177,079

Long Common Name (Study

□ Portable XR Chest AP single view	43,560
□ XR Chest Single view	25,919
□ no data	23,102
□ XR Chest AP	21,445
□ XR Chest PA and Lateral	13,517

70 more

Method (Modality)

□ XR	83,773

Showing 1 - 20 of 177,079 imaging_studies   Show Empty Columns

Browse in DICOM viewer	Study Modality	Study Description	Body Part Examine
	CR	XR CHEST 2 VIEWS	CHEST
	DX	XR CHEST 1 VW, FRONTAL	PORT CHEST
	CR	CHEST AP PORT	PORT CHEST
	CR	XR PORT CHEST 1V	
	DX	XR PORT CHEST 1V	CHEST

https://github.com/MIDRC/midrc_dicom_harmonization
https://loinc.org/kb/users-guide/loinc-rsna-radiology-playbook-user-guide/

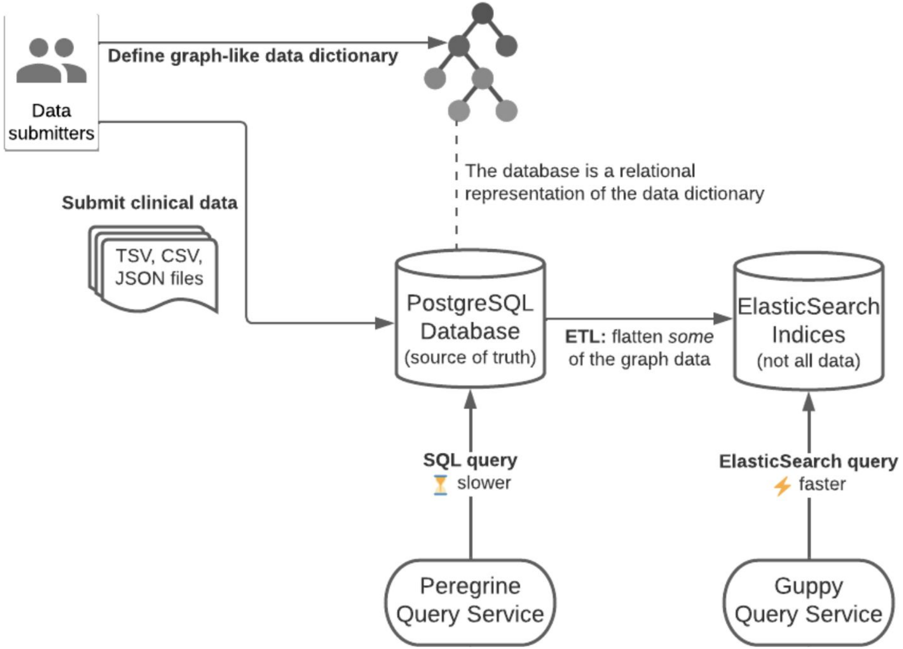
Running the ETL

Once all data have been submitted to the graph prior to a data release, the ETL process is performed.

The ETL flattens select properties from the graph into ElasticSearch indices that can be queried by guppy.

Guppy indices / queries power the data explorer GUI.

Exploration

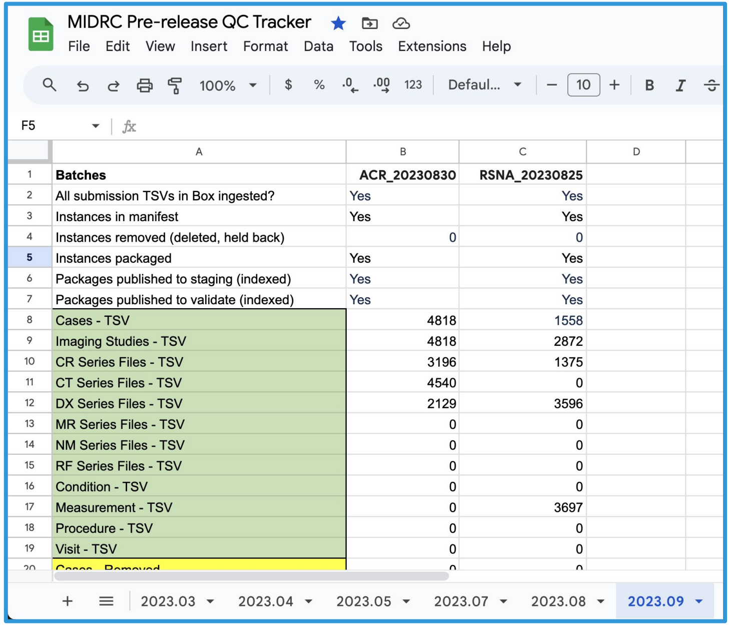

Data submitters
Define graph-like data dictionary
Submit clinical data
TSV, CSV, JSON files
The database is a relational representation of the data dictionary
PostgreSQL Database (source of truth)
ETL: flatten *some* of the graph data
ElasticSearch Indices (not all data)
SQL query
slower
ElasticSearch query
faster
Peregrine Query Service
Guppy Query Service

https://github.com/uc-cdis/tube/blob/master/docs/OVERVIEW.md

Pre-release QC

Prior to releasing new data from Staging to Production, Gen3 performs a Pre-release QC Checklist.

- Calculate derived properties / LOINC mapping.
- Confirm ETL has been run.
- Confirm counts of files and metadata entities submitted match expectations.
- Confirm UI components / file downloads working.
- Confirm data dictionary versions are up-to-date and match between staging environments.
- Confirm software versions are up-to-date and match between staging environments.
- New tutorial Jupyter notebooks are added to resource browser.



MIDRC Pre-release QC Tracker

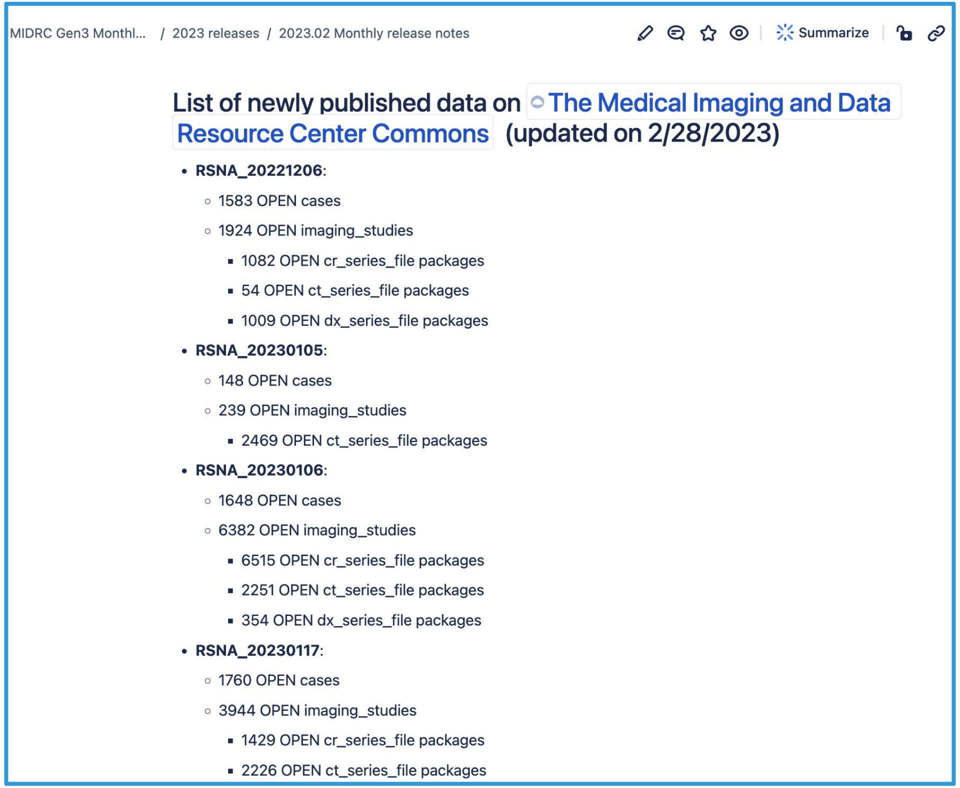
	A	B	C	D
1	**Batches**	**ACR_20230830**	**RSNA_20230825**	
2	All submission TSVs in Box ingested?	Yes	Yes	
3	Instances in manifest	Yes	Yes	
4	Instances removed (deleted, held back)	0	0	
5	Instances packaged	Yes	Yes	
6	Packages published to staging (indexed)	Yes	Yes	
7	Packages published to validate (indexed)	Yes	Yes	
8	Cases - TSV	4818	1558	
9	Imaging Studies - TSV	4818	2872	
10	CR Series Files - TSV	3196	1375	
11	CT Series Files - TSV	4540	0	
12	DX Series Files - TSV	2129	3596	
13	MR Series Files - TSV	0	0	
14	NM Series Files - TSV	0	0	
15	RF Series Files - TSV	0	0	
16	Condition - TSV	0	0	
17	Measurement - TSV	0	3697	
18	Procedure - TSV	0	0	
19	Visit - TSV	0	0	
20	Cases - Removed	0	0	

https://github.com/uc-cdis/midrc-etl/blob/master/QC/MIDRC_QC_prerelease_workflow.ipynb

Data Release / Publication Process

At the end of every month, Gen3 performs a “release”:

- Staging indexd, MDS, and sheepdog **databases are copied** from staging to production.
- Relevant **data-portal config is copied to prod** (gitops.json, manifest.json and ETL mapping).
- Finally, **ETL is run in production** to update guppy indices.



MIDRC Gen3 Monthl... / 2023 releases / 2023.02 Monthly release notes
Summarize

List of newly published data on The Medical Imaging and Data Resource Center Commons (updated on 2/28/2023)

- **RSNA_20221206**:
  - 1583 OPEN cases
  - 1924 OPEN imaging_studies
    - 1082 OPEN cr_series_file packages
    - 54 OPEN ct_series_file packages
    - 1009 OPEN dx_series_file packages
- **RSNA_20230105**:
  - 148 OPEN cases
  - 239 OPEN imaging_studies
    - 2469 OPEN ct_series_file packages
- **RSNA_20230106**:
  - 1648 OPEN cases
  - 6382 OPEN imaging_studies
    - 6515 OPEN cr_series_file packages
    - 2251 OPEN ct_series_file packages
    - 354 OPEN dx_series_file packages
- **RSNA_20230117**:
  - 1760 OPEN cases
  - 3944 OPEN imaging_studies
    - 1429 OPEN cr_series_file packages
    - 2226 OPEN ct_series_file packages

Example release notes from a MIDRC monthly release.

Thank You!

- Gen3 / Center for Translational Data Science
  - Robert Grossman (co-PI)
  - PMs
    - Ao Liu
    - Lynette Lilly
    - Karen Hyatt
    - Devin Grant-Keane
  - User Services Team
    - Johnbright Anyaibe
    - Eric Giger
    - Dan Biber
    - Tara Lichtenberg
  - Technical Leads
    - Pauline Ribeyre
    - Sai Shanmukha Narumanchi
    - Andrew Prokhorenkov
    - Thanh Nguyen

https://www.midrc.org/midrc-team

- MIDRC Central Admin / UChicago
  - Maryellen Giger (co-PI)
  - Katie Pizer (Lead Admin)
  - Erin Mueller (Lead Admin)
  - Nick Gruszauskas (HIRO)
- RSNA
  - Curtis Langlotz (co-PI)
  - Adam Flanders (co-PI)
  - Chris Carr (Data Lead)
- ACR
  - Charles Apgar (co-PI)
  - Michael Tilkin (co-PI)
  - Tao Wang (Data Lead)
  - Brian Bialecki (Data Lead)
- AAPM
  - Maryellen Giger (co-PI)
  - Paul Kinahan (co-PI)
- **And many many more!**

g3t: Gen3 Tracker – User Driven Submissions

Jordan Lee and Liam Beckman

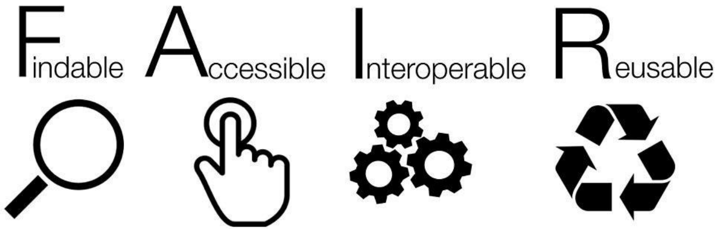
Development By: Brian Walsh, Matthew Peterkort, Nasim Sanati, and Quinn Wai Wong

Ellrott Lab, Oregon Health and Science University

"What is the biggest open challenge in biology?"

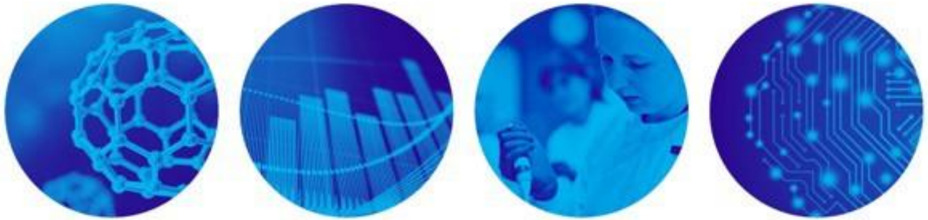
- Getting people to share data.
- Structuring, organizing, and annotating data with metadata so it's useful.
- Building higher-level abstractions so people can efficiently work with big data.

Vince Buffalo



- **Findable (F)**
  - **Metadata:** Ensure data is accompanied by rich metadata for easy discovery.
  - **Unique Identifier:** Assign a unique and persistent identifier to the dataset.
  - **Searchable:** Enhance findability through search engines and repositories.
- **Accessible (A)**
  - **Open Access:** Make data openly accessible to a wide range of users.
  - **Permissions:** Clearly define access rights and provide necessary permissions.
  - **Formats:** Ensure data is available in multiple formats for different user needs.
- **Interoperable (I)**
  - **Standards:** Use common data standards and formats to facilitate interoperability.
  - **Linkage:** Enable linkage with other datasets to derive additional insights.
  - **APIs:** Provide Application Programming Interfaces (APIs) for seamless integration.
- **Reusable (R)**
  - **Documentation:** Provide comprehensive documentation for easy understanding.
  - **Licenses:** Clearly specify the terms of use and licensing agreements.
  - **Citations:** Encourage and facilitate proper citation for data reuse.

ACED INTERNATIONAL ALLIANCE FOR CANCER EARLY DETECTION

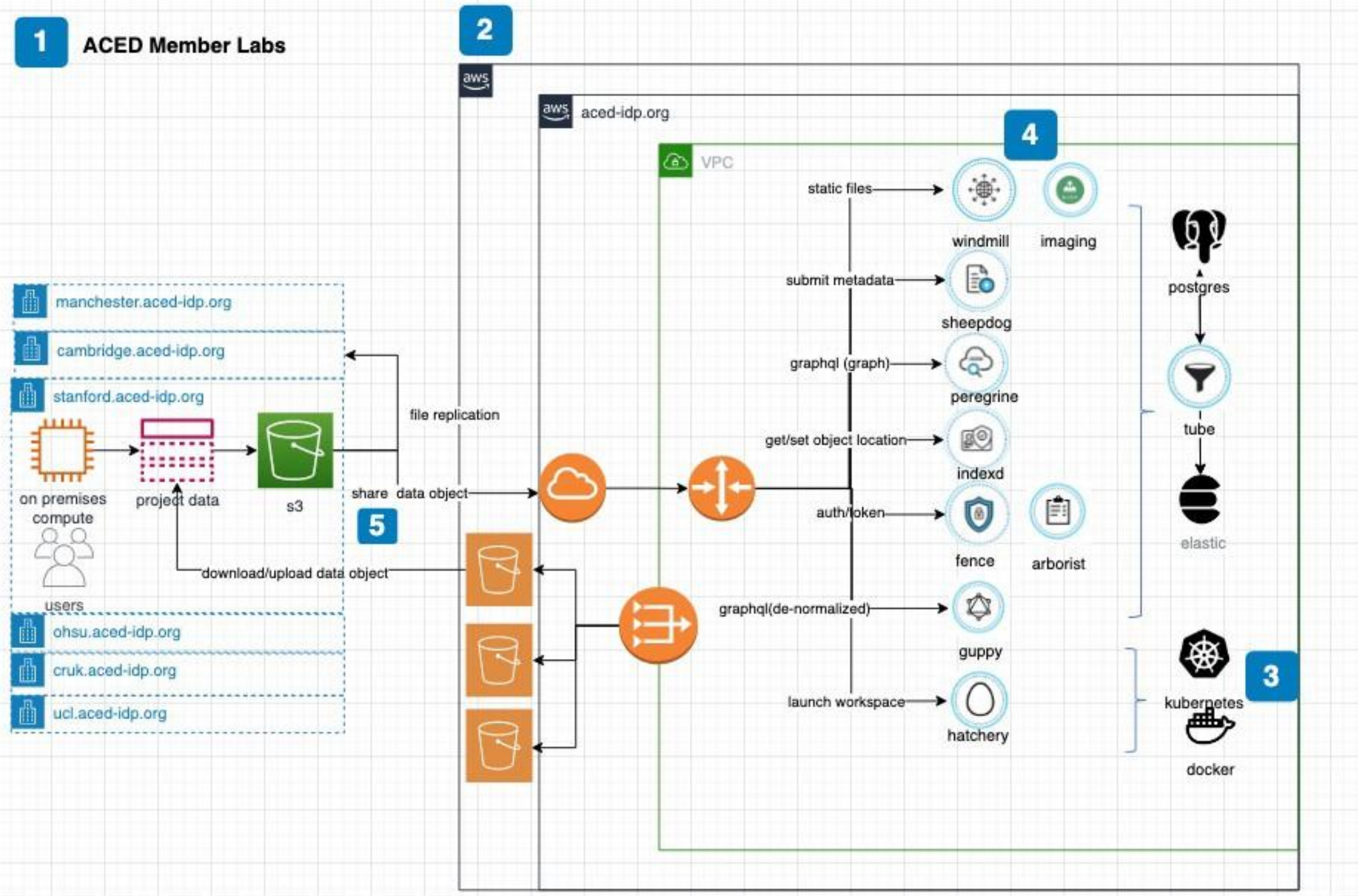


We are uniting world leading researchers to tackle the biggest challenges in early detection, an important area of unmet clinical need. Scientists in the Alliance are working together at the forefront of technological innovation to translate research into realistic ways to **improve cancer diagnosis**, which can be **implemented into health systems** and meaningfully benefit people with cancer.

cancerresearchuk.org

ACED IDP

High Level Architecture



1 ACED Member Labs

2

aws

aws aced-idp.org

VPC

4

manchester.aced-idp.org

cambridge.aced-idp.org

stanford.aced-idp.org

on premises compute

project data

s3

users

ohsu.aced-idp.org

cruk.aced-idp.org

ucl.aced-idp.org

file replication

share data object

5

download/upload data object

static files

submit metadata

graphql (graph)

get/set object location

auth/token

graphql(de-normalized)

launch workspace

windmill

imaging

sheepdog

peregrine

indexd

fence

arborist

guppy

hatchery

postgres

tube

elastic

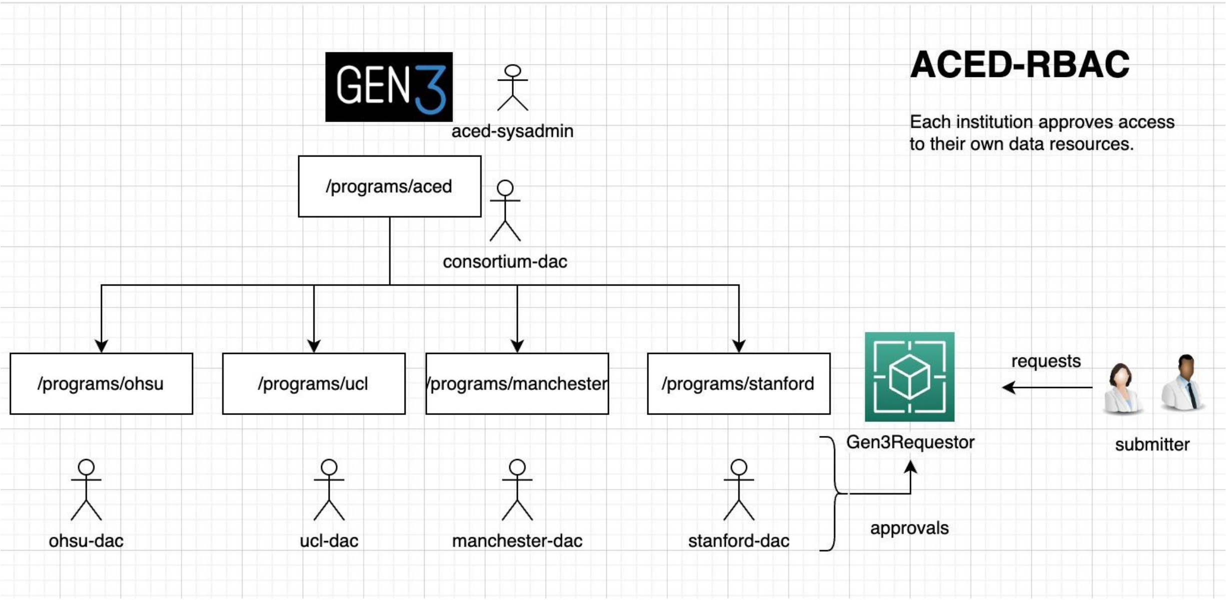
kubernetes

3

docker

Data Stewardship

As an aced data steward, in order to understand my role in creating projects and granting access, I need way to understand and implement my role and responsibilities.



ACED-RBAC

Each institution approves access to their own data resources.

Accessible

Open Access, Permissions : A distributed team controls data stewardship to grant and revoke access over their institution’s data.

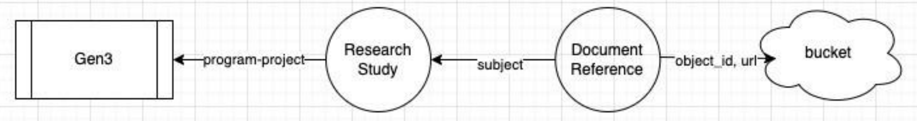
aced-rbac

Data Stewardship: Project Creation

```
# as a data submitter
g3t init {program}-{project}

# as a data steward
g3t collaborator approve --request_id {request ID}

# as a system administrator
g3t projects create
```

Minimal Viable Study

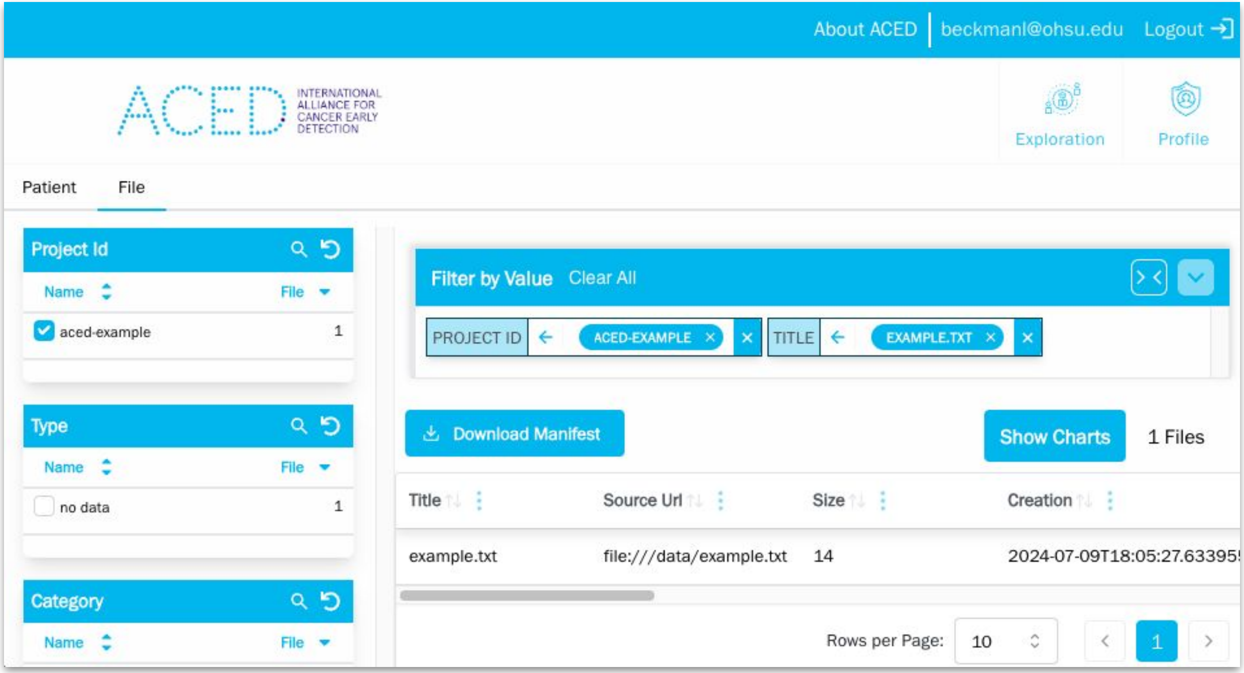
As a data submitter, in order to share data, I want to upload a set of files



```
# repeat for each file
g3t add PATH [--size,--<hash>,--mime]

# create metadata
g3t meta init

# add to repository
g3t commit –m “My study’s files”
g3t push

# view upload status (pending, complete)
g3t status
```

Minimal Viable Study

```
$ ls -1 META/

DocumentReference.ndjson
ResearchStudy.ndjson
```



Project Id

Name	File
☑ aced-example	1

Type

Name	File
□ no data	1

Category

Name	File

Filter by Value Clear All

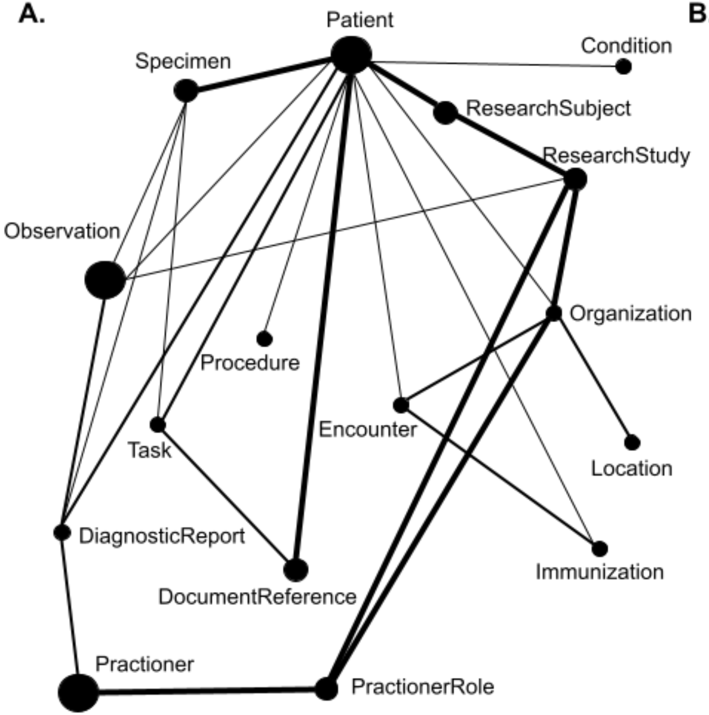
PROJECT ID ← ACED-EXAMPLE TITLE ← EXAMPLE.TXT

1 Files

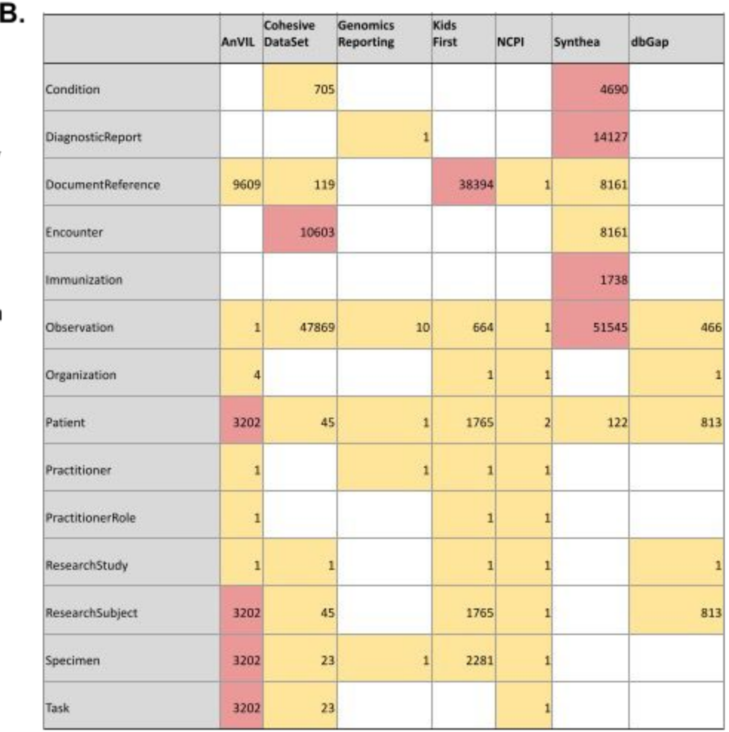
Title	Source Url	Size	Creation
example.txt	file:///data/example.txt	14	2024-07-09T18:05:27.63395

Rows per Page: 10 1

Data Model



B.

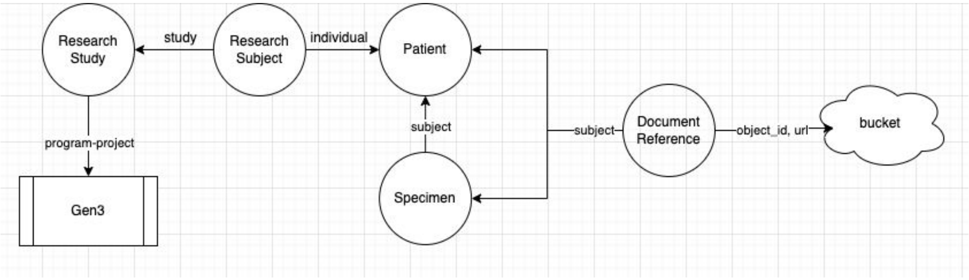


	AnVIL	Cohesive DataSet	Genomics Reporting	Kids First	NCPI	Synthea	dbGap
Condition		705				4690	
DiagnosticReport			1			14127	
DocumentReference	9609	119		38394	1	8161	
Encounter		10603				8161	
Immunization						1738	
Observation	1	47869	10	664	1	51545	466
Organization	4			1	1		1
Patient	3202	45	1	1765	2	122	813
Practitioner	1		1	1	1		
PractitionerRole	1			1	1		
ResearchStudy	1	1		1	1		1
ResearchSubject	3202	45		1765	1		813
Specimen	3202	23	1	2281	1		
Task	3202	23			1		

Aggregated demonstration datasets. A) A Graph representation of the unified schema, with vertex sizes representing the relative number of data sets that contain that vertex type and the thickness of the edges representing the number of datasets that implemented that relationship. B) A table of the total vertex type counts across the reported datasets.

Study with Tagged Patients [specimens, etc]

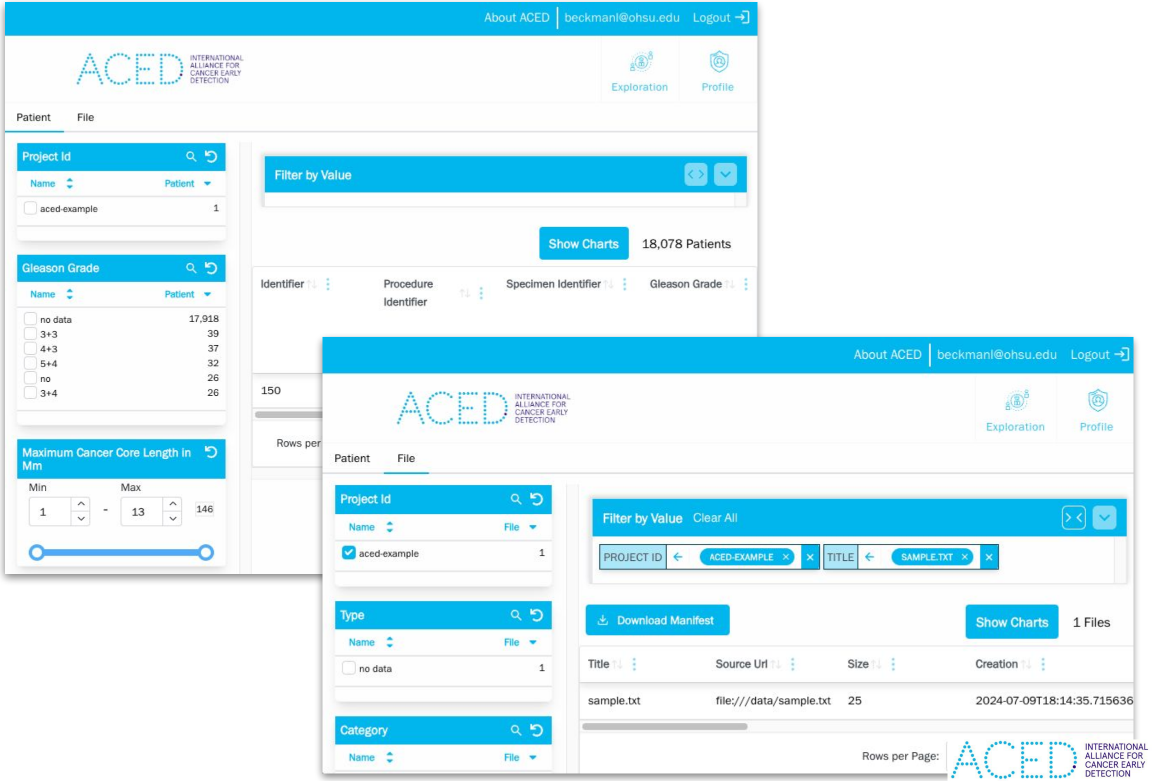
As a data submitter, in order to share data, I want to upload a set of files, each tagged with any of patient, specimen, task, etc.



```
# for each file: …
g3t add PATH --patient my-patient-identifier [--specimen, --size, --hash <hash>]

# create metadata
g3t utilities meta create
# optional: edit generated metadata

# add to repository
g3t commit –m “my study’s files, subjects and/or specimens”
g3t push
```

Study with Tagged Patients [specimens, etc]

```
$ ls -1 META/

DocumentReference.ndjson
Patient.ndjson
ResearchStudy.ndjson
ResearchSubject.ndjson
```



Project Id

Name	Patient
□ aced-example	1

Gleason Grade

Name	Patient
□ no data	17,918
□ 3+3	39
□ 4+3	37
□ 5+4	32
□ no	26
□ 3+4	26

Maximum Cancer Core Length in Mm

Min 1 - Max 13 146

18,078 Patients

Identifier	Procedure Identifier	Specimen Identifier	Gleason Grade

150

Rows per

Project Id

Name	File
☑ aced-example	1

Type

Name	File
□ no data	1

Category

Name	File

Filter by Value Clear All

PROJECT ID ← ACED-EXAMPLE × | TITLE ← SAMPLE.TXT ×

1 Files

Title	Source Url	Size	Creation
sample.txt	file:///data/sample.txt	25	2024-07-09T18:14:35.715636

Rows per Page:

Deep Dive

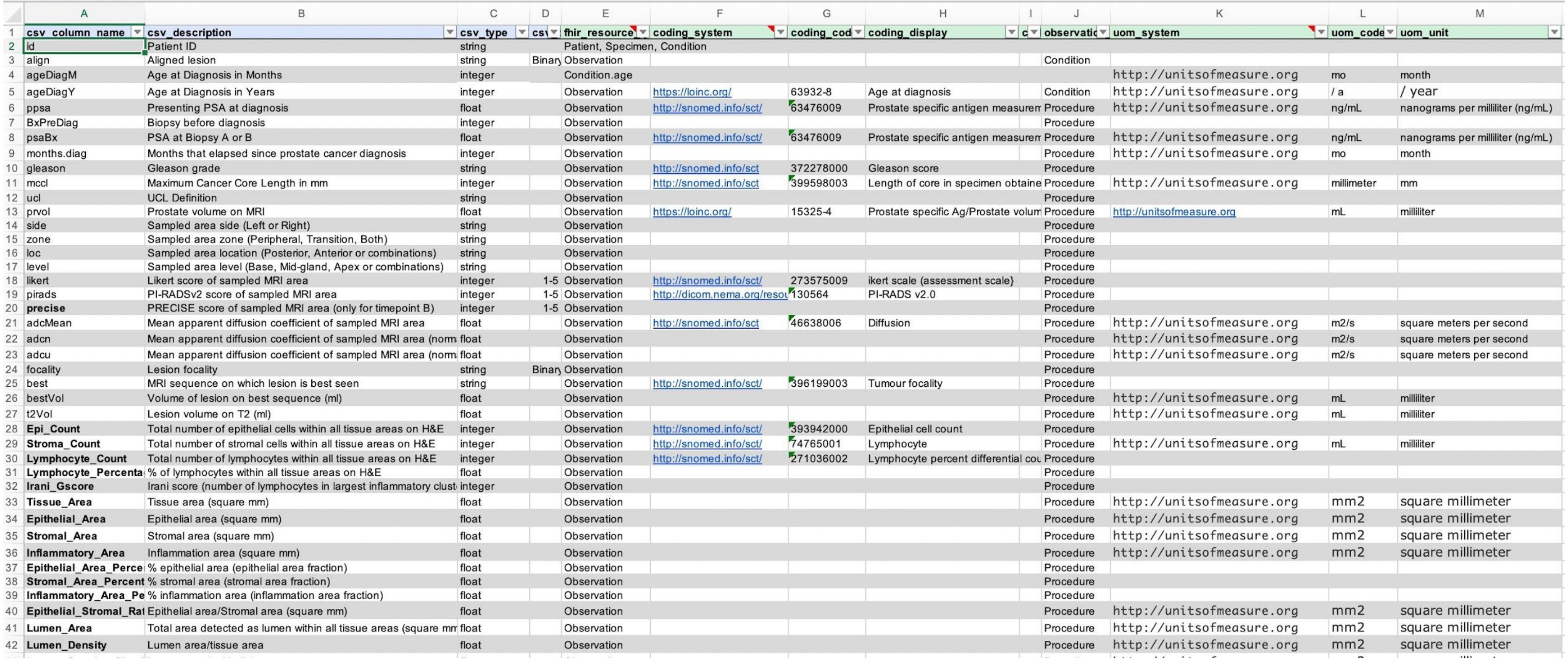
```
{
  "resourceType": "Patient",
  "id": "f027d9b9-da61-5f48-9378-f4dc0e6b85e6",
  "identifier": [
      {
      "use": "official",
      "system": "https://aced-idp.org/test-one_patient",
      "value": "P1"
      }
  ]
}
```

Findable

- **Findable (F)**
- **Metadata:** The aced-idp system encourages and facilitates the creation of metadata over a wide variety of use cases
- **Unique Identifier:** The system requires and maintains a submitter driven identifier and well as location independent, idempotent ids for all metadata resources. File objects are also registered as DRS (GAGH Data Repository
Service) uris
- **Searchable:** All of the above keys are searchable via the portal or API. The system defaults CodeableConcept attributes to submitter provided values and encourages additional tagging with standard ontology terms

Study with Rich Set of Measurements

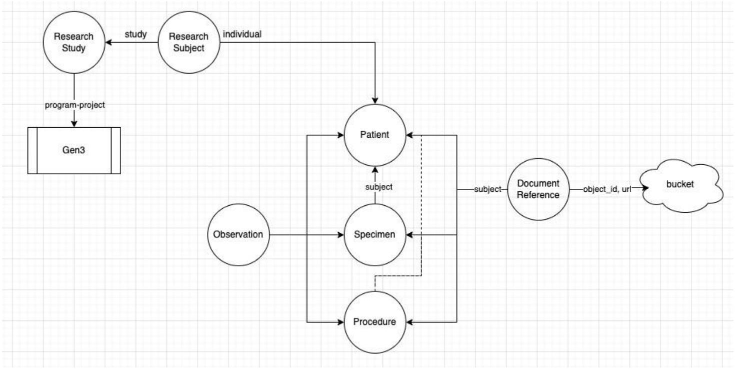
Transform submission CSVs to FHIR



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	**csv_column_name**	**csv_description**	**csv_type**	**csv**	**fhir_resource**	**coding_system**	**coding_cod**	**coding_display**	**c**	**observatic**	**uom_system**	**uom_code**	**uom_unit**
2	id	Patient ID	string		Patient, Specimen, Condition								
3	align	Aligned lesion	string	Binary	Observation					Condition			
4	ageDiagM	Age at Diagnosis in Months	integer		Condition.age						http://unitsofmeasure.org	mo	month
5	ageDiagY	Age at Diagnosis in Years	integer		Observation	https://loinc.org/	63932-8	Age at diagnosis		Condition	http://unitsofmeasure.org	/ a	/ year
6	ppsa	Presenting PSA at diagnosis	float		Observation	http://snomed.info/sct/	63476009	Prostate specific antigen measurem		Procedure	http://unitsofmeasure.org	ng/mL	nanograms per milliliter (ng/mL)
7	BxPreDiag	Biopsy before diagnosis	integer		Observation					Procedure			
8	psaBx	PSA at Biopsy A or B	float		Observation	http://snomed.info/sct/	63476009	Prostate specific antigen measurem		Procedure	http://unitsofmeasure.org	ng/mL	nanograms per milliliter (ng/mL)
9	months.diag	Months that elapsed since prostate cancer diagnosis	integer		Observation					Procedure	http://unitsofmeasure.org	mo	month
10	gleason	Gleason grade	string		Observation	http://snomed.info/sct	372278000	Gleason score		Procedure			
11	mccl	Maximum Cancer Core Length in mm	integer		Observation	http://snomed.info/sct	399598003	Length of core in specimen obtaine		Procedure	http://unitsofmeasure.org	millimeter	mm
12	ucl	UCL Definition	string		Observation					Procedure			
13	prvol	Prostate volume on MRI	float		Observation	https://loinc.org/	15325-4	Prostate specific Ag/Prostate volum		Procedure	http://unitsofmeasure.org	mL	milliliter
14	side	Sampled area side (Left or Right)	string		Observation					Procedure			
15	zone	Sampled area zone (Peripheral, Transition, Both)	string		Observation					Procedure			
16	loc	Sampled area location (Posterior, Anterior or combinations)	string		Observation					Procedure			
17	level	Sampled area level (Base, Mid-gland, Apex or combinations)	string		Observation					Procedure			
18	likert	Likert score of sampled MRI area	integer	1-5	Observation	http://snomed.info/sct/	273575009	ikert scale (assessment scale}		Procedure			
19	pirads	PI-RADSv2 score of sampled MRI area	integer	1-5	Observation	http://dicom.nema.org/resou	130564	PI-RADS v2.0		Procedure			
20	**precise**	PRECISE score of sampled MRI area (only for timepoint B)	integer	1-5	Observation					Procedure			
21	adcMean	Mean apparent diffusion coefficient of sampled MRI area	float		Observation	http://snomed.info/sct	46638006	Diffusion		Procedure	http://unitsofmeasure.org	m2/s	square meters per second
22	adcn	Mean apparent diffusion coefficient of sampled MRI area (norm	float		Observation					Procedure	http://unitsofmeasure.org	m2/s	square meters per second
23	adcu	Mean apparent diffusion coefficient of sampled MRI area (norm	float		Observation					Procedure	http://unitsofmeasure.org	m2/s	square meters per second
24	focality	Lesion focality	string	Binary	Observation					Procedure			
25	best	MRI sequence on which lesion is best seen	string		Observation	http://snomed.info/sct/	396199003	Tumour focality		Procedure			
26	bestVol	Volume of lesion on best sequence (ml)	float		Observation					Procedure	http://unitsofmeasure.org	mL	milliliter
27	t2Vol	Lesion volume on T2 (ml)	float		Observation					Procedure	http://unitsofmeasure.org	mL	milliliter
28	**Epi_Count**	Total number of epithelial cells within all tissue areas on H&E	integer		Observation	http://snomed.info/sct/	393942000	Epithelial cell count		Procedure			
29	**Stroma_Count**	Total number of stromal cells within all tissue areas on H&E	integer		Observation	http://snomed.info/sct/	74765001	Lymphocyte		Procedure	http://unitsofmeasure.org	mL	milliliter
30	**Lymphocyte_Count**	Total number of lymphocytes within all tissue areas on H&E	integer		Observation	http://snomed.info/sct/	271036002	Lymphocyte percent differential cou		Procedure			
31	**Lymphocyte_Percenta**	% of lymphocytes within all tissue areas on H&E	float		Observation					Procedure			
32	**Irani_Gscore**	Irani score (number of lymphocytes in largest inflammatory clust	integer		Observation					Procedure			
33	**Tissue_Area**	Tissue area (square mm)	float		Observation					Procedure	http://unitsofmeasure.org	mm2	square millimeter
34	**Epithelial_Area**	Epithelial area (square mm)	float		Observation					Procedure	http://unitsofmeasure.org	mm2	square millimeter
35	**Stromal_Area**	Stromal area (square mm)	float		Observation					Procedure	http://unitsofmeasure.org	mm2	square millimeter
36	**Inflammatory_Area**	Inflammation area (square mm)	float		Observation					Procedure	http://unitsofmeasure.org	mm2	square millimeter
37	**Epithelial_Area_Perce**	% epithelial area (epithelial area fraction)	float		Observation					Procedure			
38	**Stromal_Area_Percent**	% stromal area (stromal area fraction)	float		Observation					Procedure			
39	**Inflammatory_Area_Pe**	% inflammation area (inflammation area fraction)	float		Observation					Procedure			
40	**Epithelial_Stromal_Rat**	Epithelial area/Stromal area (square mm)	float		Observation					Procedure	http://unitsofmeasure.org	mm2	square millimeter
41	**Lumen_Area**	Total area detected as lumen within all tissue areas (square mm	float		Observation					Procedure	http://unitsofmeasure.org	mm2	square millimeter
42	**Lumen_Density**	Lumen area/tissue area	float		Observation					Procedure	http://unitsofmeasure.org	mm2	square millimeter

Study with Rich Set of Measurements

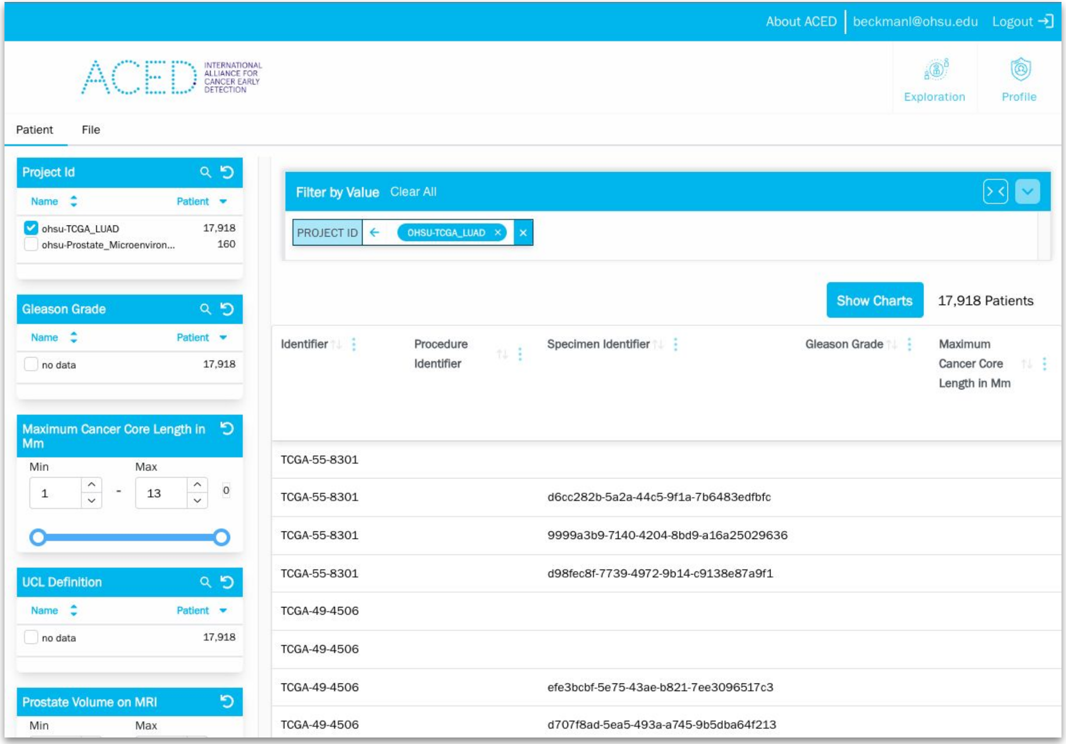
As a data submitter, in order to share data, I want to upload a set of files accompanied with a rich set of observations



```
# for each file: …
g3t add PATH --patient my-patient-identifier [--specimen --size, --hash <hash>]

# create metadata using a transformer
G3T_PLUGIN=my_project.transformer g3t_etl transform

#add to repository
g3t commit –m “my study’s files, subjects and/or specimens”
g3t push
```

Study with Rich Set of Measurements

```
$ ls -l META/

Condition.ndjson
Observation.ndjson
Patient.ndjson
Procedure.ndjson
ResearchStudy.ndjson
ResearchSubject.ndjson
```



Project Id

Name	Patient
☑ ohsu-TCGA_LUAD	17,918
□ ohsu-Prostate_Microenviron...	160

Gleason Grade

Name	Patient
□ no data	17,918

Maximum Cancer Core Length in Mm

Min 1 - Max 13

UCL Definition

Name	Patient
□ no data	17,918

Prostate Volume on MRI

Min Max

PROJECT ID ← OHSU-TCGA_LUAD ×

17,918 Patients

Identifier	Procedure Identifier	Specimen Identifier	Gleason Grade	Maximum Cancer Core Length in Mm
TCGA-55-8301				
TCGA-55-8301		d6cc282b-5a2a-44c5-9f1a-7b6483edfbfc		
TCGA-55-8301		9999a3b9-7140-4204-8bd9-a16a25029636		
TCGA-55-8301		d98fec8f-7739-4972-9b14-c9138e87a9f1		
TCGA-49-4506				
TCGA-49-4506				
TCGA-49-4506		efe3bcbf-5e75-43ae-b821-7ee3096517c3		
TCGA-49-4506		d707f8ad-5ea5-493a-a745-9b5dba64f213		

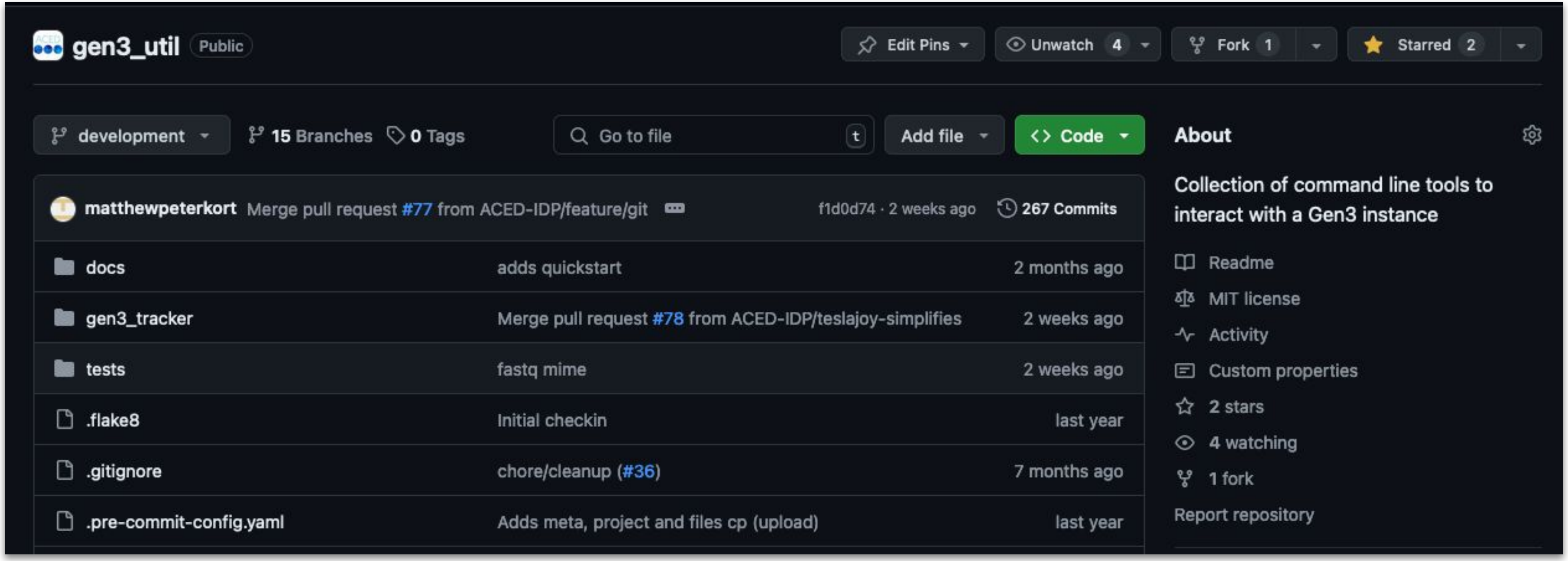
Next Steps

Potential Improvements to data upload and `g3t` on our Roadmap (suggestions welcome!):

- Add ability to upload multiple files in parallel (either as an entire directory or other specified set)
- User Friendliness
  - Expand documentation and overall UX based on data analysts experiences
  - Add lessons/tutorials for easier and more gradual adoption
- Learning and Sharing with the Gen3 Community
  - Alternative ways to manage Gen3 data, different Use Cases
- Continue integration with gen3-client + Frontend Framework

Development + Contributions

g3t itself is hosted on a public repo (with a Contributor guide) — Issues + PR’s welcome!



matthewpeterkort Merge pull request #77 from ACED-IDP/feature/git f1d0d74 · 2 weeks ago 267 Commits

docs	adds quickstart	2 months ago
gen3_tracker	Merge pull request #78 from ACED-IDP/teslajoy-simplifies	2 weeks ago
tests	fastq mime	2 weeks ago
.flake8	Initial checkin	last year
.gitignore	chore/cleanup (#36)	7 months ago
.pre-commit-config.yaml	Adds meta, project and files cp (upload)	last year

About

Collection of command line tools to interact with a Gen3 instance

- Readme
- MIT license
- Activity
- Custom properties
- 2 stars
- 4 watching
- 1 fork

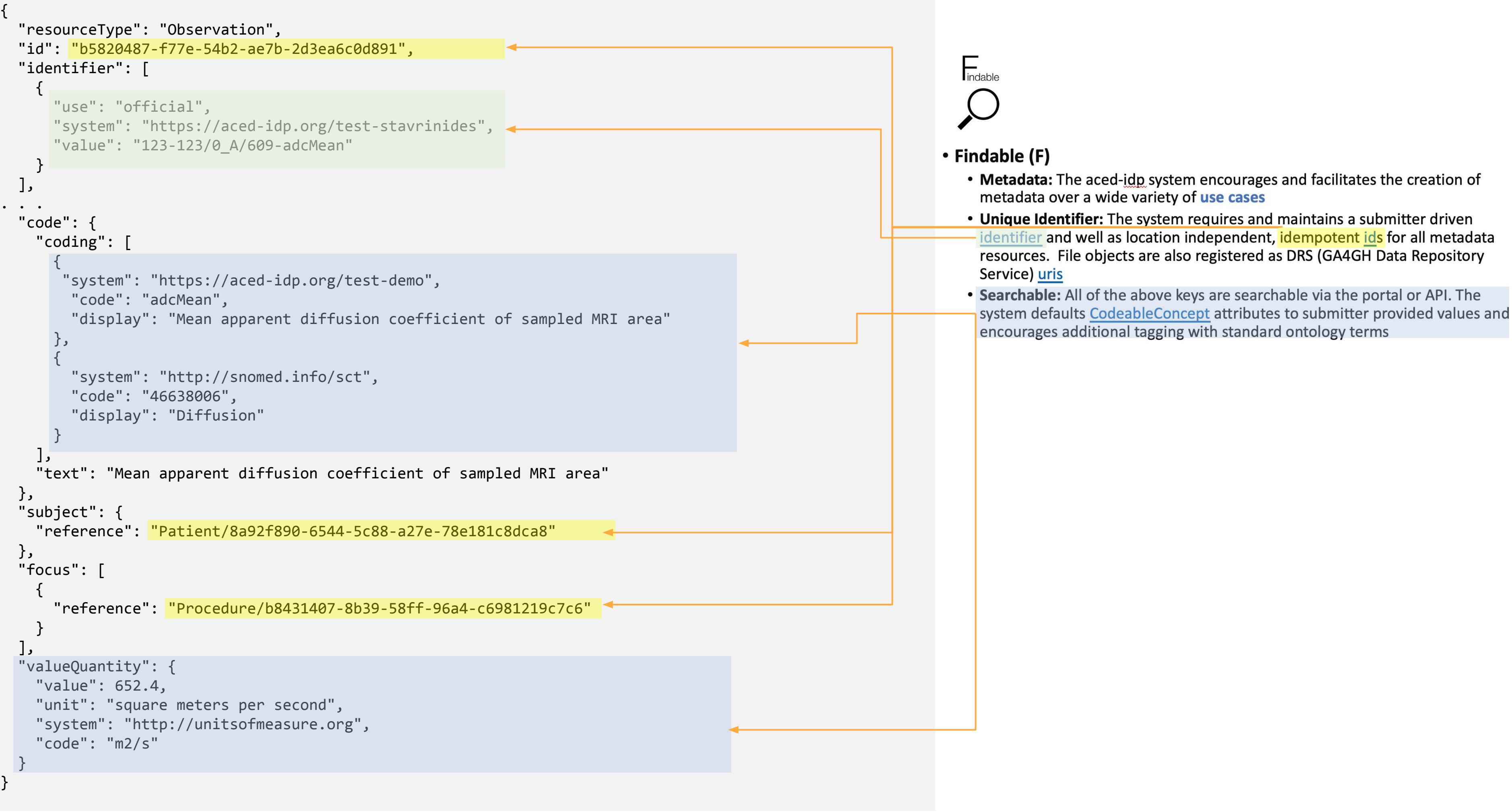
Report repository

ANY
QUESTIONS?

schema management

```
# This limits the top level objects the system will
render dependency_order:
  # gen3 scaffolding required objects
  - _definitions.yaml
  - _terms.yaml
  - Program
  - Project
  # FHIR objects
  - Organization
  - Practitioner
  - PractitionerRole
  - ResearchStudy
  - Patient
  - ResearchSubject
  - Substance
  - Specimen
  - Observation
  - DiagnosticReport
  - Condition
  - Medication
  - MedicationAdministration
  - Procedure
  - DocumentReference
  - Task
  - ImagingStudy
  - FamilyMemberHistory
  - BodyStructure
```

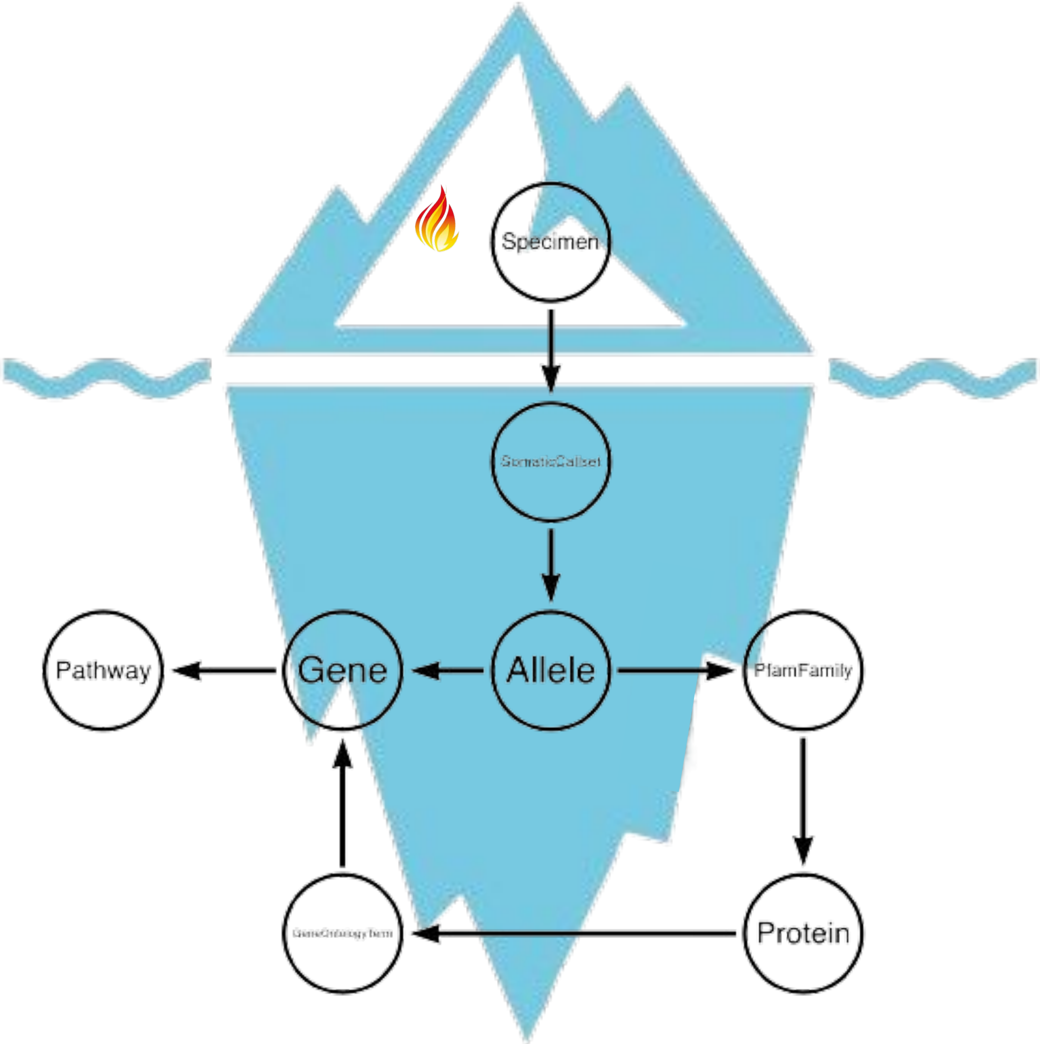
The iceberg schema tools project enables the developer to manage schema “scope” and link to research entities.



```
{
  "resourceType": "Observation",
  "id": "b5820487-f77e-54b2-ae7b-2d3ea6c0d891",
  "identifier": [
    {
      "use": "official",
      "system": "https://aced-idp.org/test-stavrinides",
      "value": "123-123/0_A/609-adcMean"
    }
  ],
. . .
  "code": {
    "coding": [
      {
       "system": "https://aced-idp.org/test-demo",
        "code": "adcMean",
        "display": "Mean apparent diffusion coefficient of sampled MRI area"
      },
      {
        "system": "http://snomed.info/sct",
        "code": "46638006",
        "display": "Diffusion"
      }
    ],
    "text": "Mean apparent diffusion coefficient of sampled MRI area"
  },
  "subject": {
    "reference": "Patient/8a92f890-6544-5c88-a27e-78e181c8dca8"
  },
  "focus": [
    {
      "reference": "Procedure/b8431407-8b39-58ff-96a4-c6981219c7c6"
    }
  ],
  "valueQuantity": {
    "value": 652.4,
    "unit": "square meters per second",
    "system": "http://unitsofmeasure.org",
    "code": "m2/s"
  }
}
```

- **Findable (F)**
  - **Metadata:** The aced-idp system encourages and facilitates the creation of metadata over a wide variety of **use cases**
  - **Unique Identifier:** The system requires and maintains a submitter driven identifier and well as location independent, idempotent ids for all metadata resources. File objects are also registered as DRS (GA4GH Data Repository Service) uris
  - **Searchable:** All of the above keys are searchable via the portal or API. The system defaults CodeableConcept attributes to submitter provided values and encourages additional tagging with standard ontology terms

schema scope



Example: Installing g3t

g3t releases are hosted on PyPi and can be installed with your Python package manager of choice!

```
# (Optional) Set up virtual environment
python3 -m venv venv && source venv/bin/activate

# Install latest version
pip install gen3-tracker==0.0.4rc40

g3t --version
g3t, version 0.0.4rc40

export G3T_PROFILE=aced
g3t ping
msg: 'Configuration OK: Connected using profile:production'
endpoint: https://aced-idp.org
username: user@ohsu.edu
```

Example: Uploading Files

Adapted from the ACED Quickstart Guide

```
# Initialize a new project
g3t init aced-example

# Add files
g3t add folder/file.tsv
g3t add folder/file2.tsv

# Create metadata
g3t utilities meta create

# Commit files
g3t commit -m "Adding files"

# Push to the Gen3 System
g3t push
```

Example: Downloading Files

Adapted from the ACED Quickstart Guide

gen3-client is used to download files from our Gen3 system:

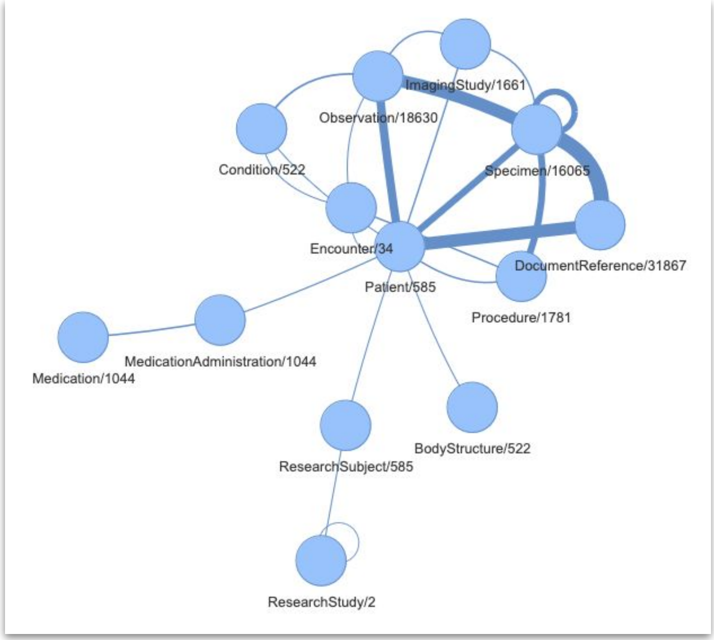
```
# Single file download via GUID
gen3-client download-single --profile=aced --guid=f623df8f-5dad-5bce-a8ca-a7b69b7805a5

# Multiple file download via file manifest
gen3-client download-multiple --profile=aced --manifest=file-manifest.json
```

Example: Utilities

Adapted from the ACED Quickstart Guide

g3t includes commands to clone projects, manage access, and view + validate metadata:

```
 # Clone an existing project
 g3t clone ohsu-TCGA_LUAD

 # View metadata as a graph
 g3t meta graph

 # Validate metadata
 g3t meta validate
 {'summary': {'DocumentReference': 31867,
'Procedure': 1781, 'Specimen': 16065, 'Medication':
1044, 'Observation': 18630, … 'Patient': 585}}
```



Data Modelling in Gen3

Joshua Harris, PhD - Research Data Manager
Australian BioCommons

Acknowledgement of Country

I would like to show my respect and Acknowledge the Traditional Custodians of the Land, of Elders past and present, on which this meeting takes place.

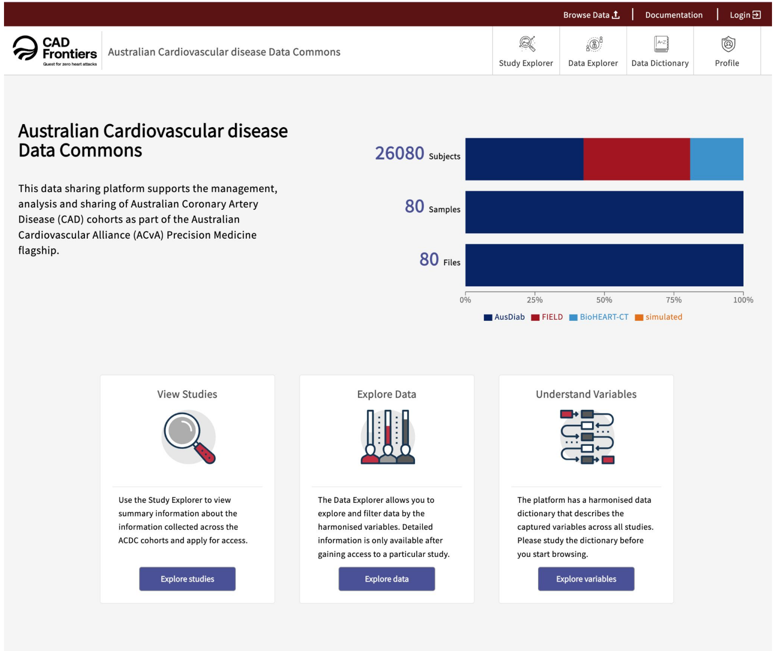
Australian BioCommons

Mission

To sustain **strategic leadership** in bioinformatics and bioscience **data infrastructure** nationally, support life science research with advanced **digital infrastructure**, provide sophisticated **analysis services**, ensure enduring access to essential digital tools, and offer comprehensive bioinformatics **training and support**.

Australian Cardiovascular disease Data Commons

GEN3 DATA COMMONS



Australian Cardiovascular disease Data Commons

This data sharing platform supports the management, analysis and sharing of Australian Coronary Artery Disease (CAD) cohorts as part of the Australian Cardiovascular Alliance (ACvA) Precision Medicine flagship.

View Studies

Use the Study Explorer to view summary information about the information collected across the ACDC cohorts and apply for access.

Explore Data

The Data Explorer allows you to explore and filter data by the harmonised variables. Detailed information is only available after gaining access to a particular study.

Understand Variables

The platform has a harmonised data dictionary that describes the captured variables across all studies. Please study the dictionary before you start browsing.

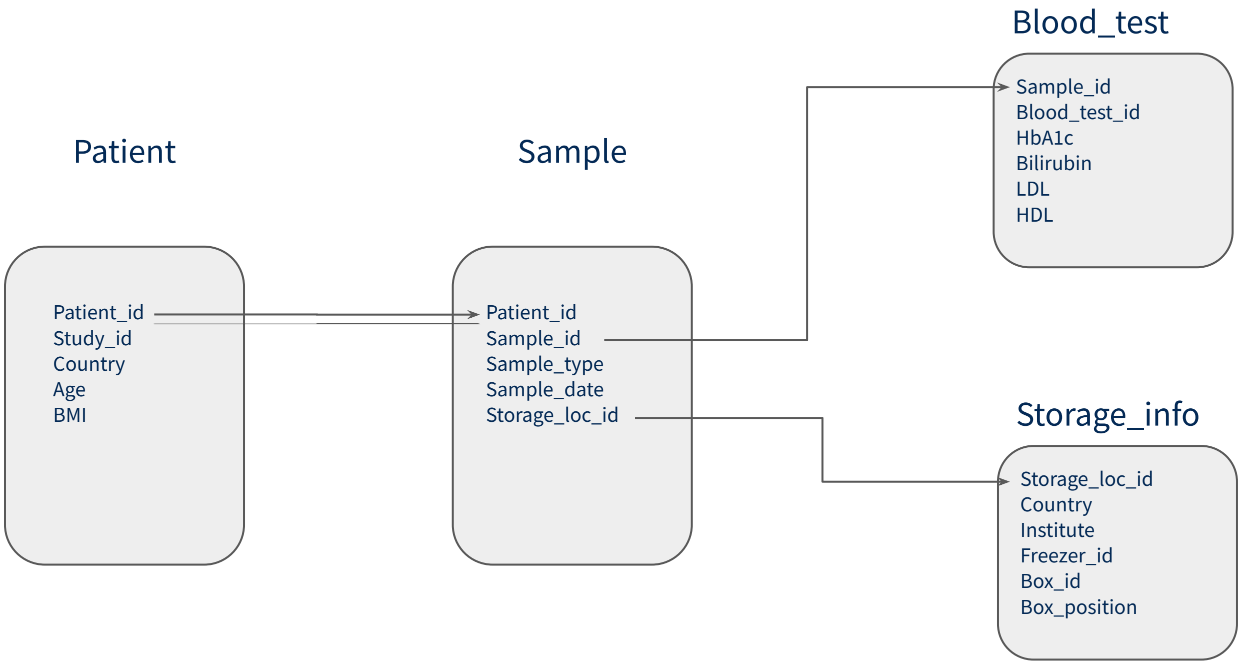
Table 1. Cohorts with available data and profiling.

Study	Outcomes (follow-up)	Total numbers[2]	Available data[1]		
			Genomic	Lipidomic	other biomarkers
AusDiab	CVD (>15 yr)	11000	0	10000	Yes
FIELD	CVD (>10 yr)	10000	5000	5000	Yes
BioHEART-CT	CTCA/CVD (<3 yr)	5000	2000	2000	proteomic, metabolomic
Busselton	CVD (>20 yr)	4492	4492	4492	WGS on 1,000
ASPREE	CVD (~5 yr)	14000	14000	4000	WGS on 2,000, Yes
LIPID	CVD (>20 yr)	10000	0	6000	Yes
45 and UP	CVD	267000	5000		WGS on 2,000
BioHEART-MI	CVD (<3 yr)	2000	2000	2000	proteomic, metabolomic
MCCS	CVD (>20 yr)	41513	12105	3000	
Baker Biobank	CVD (>15 yr)	6000	6000	0	
Caught-CAD	CTCA/CVD (<3 yr)	1000	1000	1000	
EDCAD-PMS	CTCA/CVD (<3 yr)	1000	1000	1000	
PREDICT	CVD	2500	0	0	
CDAH	CVD (>20 yr)	4947	0	0	Yes, metabolomics, imaging
ADVANCE	CVD (<5 yr)	11140	0	3779	
PROPHECY (Indigenous)	CVD (<3 yr)	1386	1386	0	proteomic, metabolomic, epigenetic
BIRCH (Indigenous)	CVD (<3 yr)	490	0	466	
DaVinci	CVD (<3 yr	600	600		
Total		**394068**	**54583**	**44737**	

[1] Available data (including ongoing profiling activities to be completed by December 2022)
[2] Represents total numbers of participants for which some (but not all) data is available.

Gen3 data modelling background

Conceptual Entity Relationship Graph



Blood_test

Sample_id
Blood_test_id
HbA1c
Bilirubin
LDL
HDL

Patient

Patient_id
Study_id
Country
Age
BMI

Sample

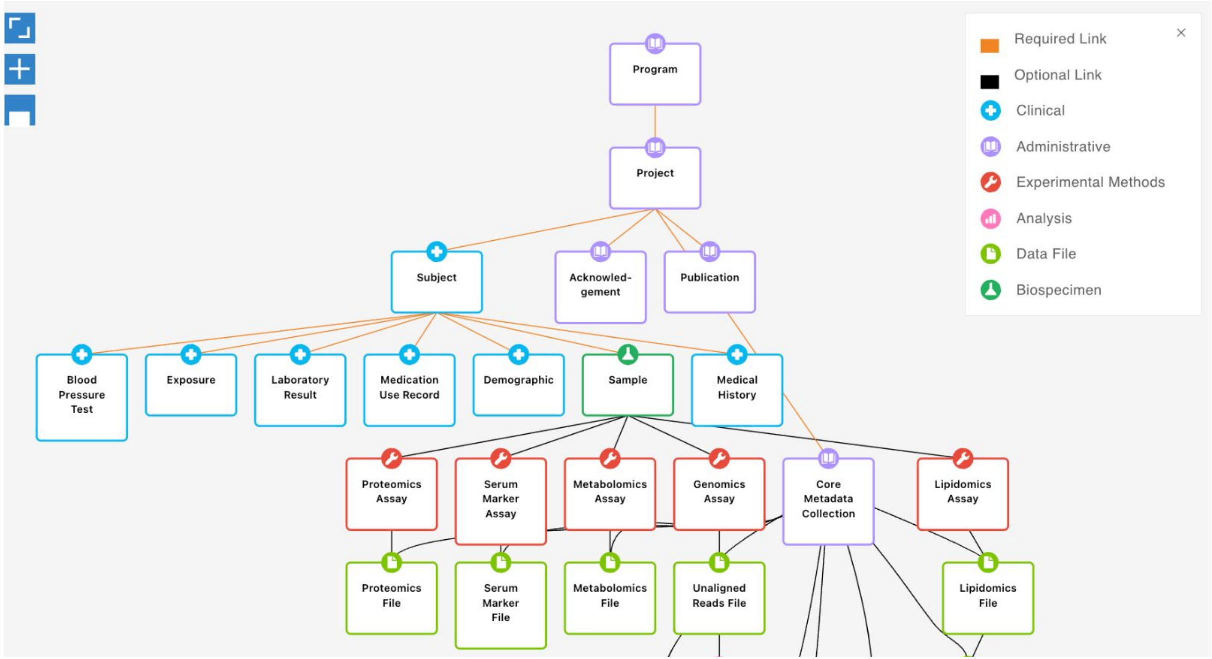
Patient_id
Sample_id
Sample_type
Sample_date
Storage_loc_id

Storage_info

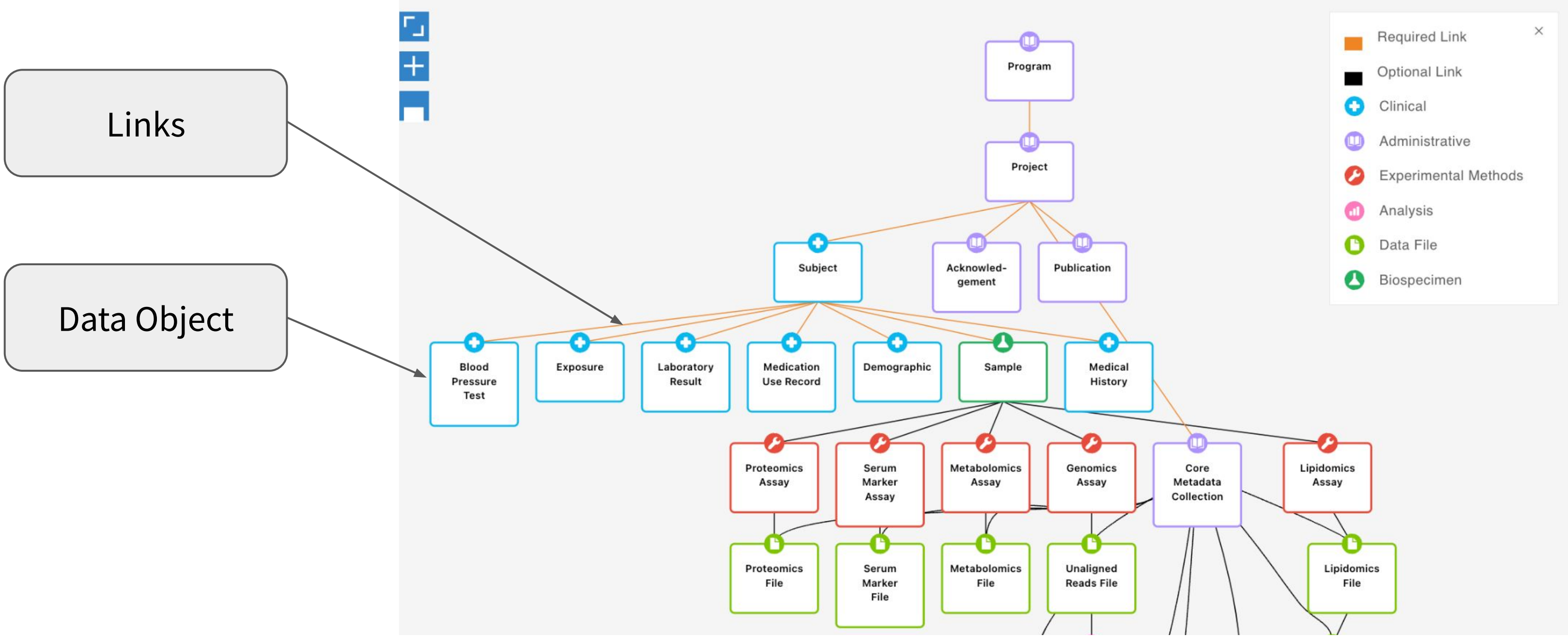
Storage_loc_id
Country
Institute
Freezer_id
Box_id
Box_position

Gen3 data modelling background

Gen3's Graph View Provides a conceptual overview of a data model



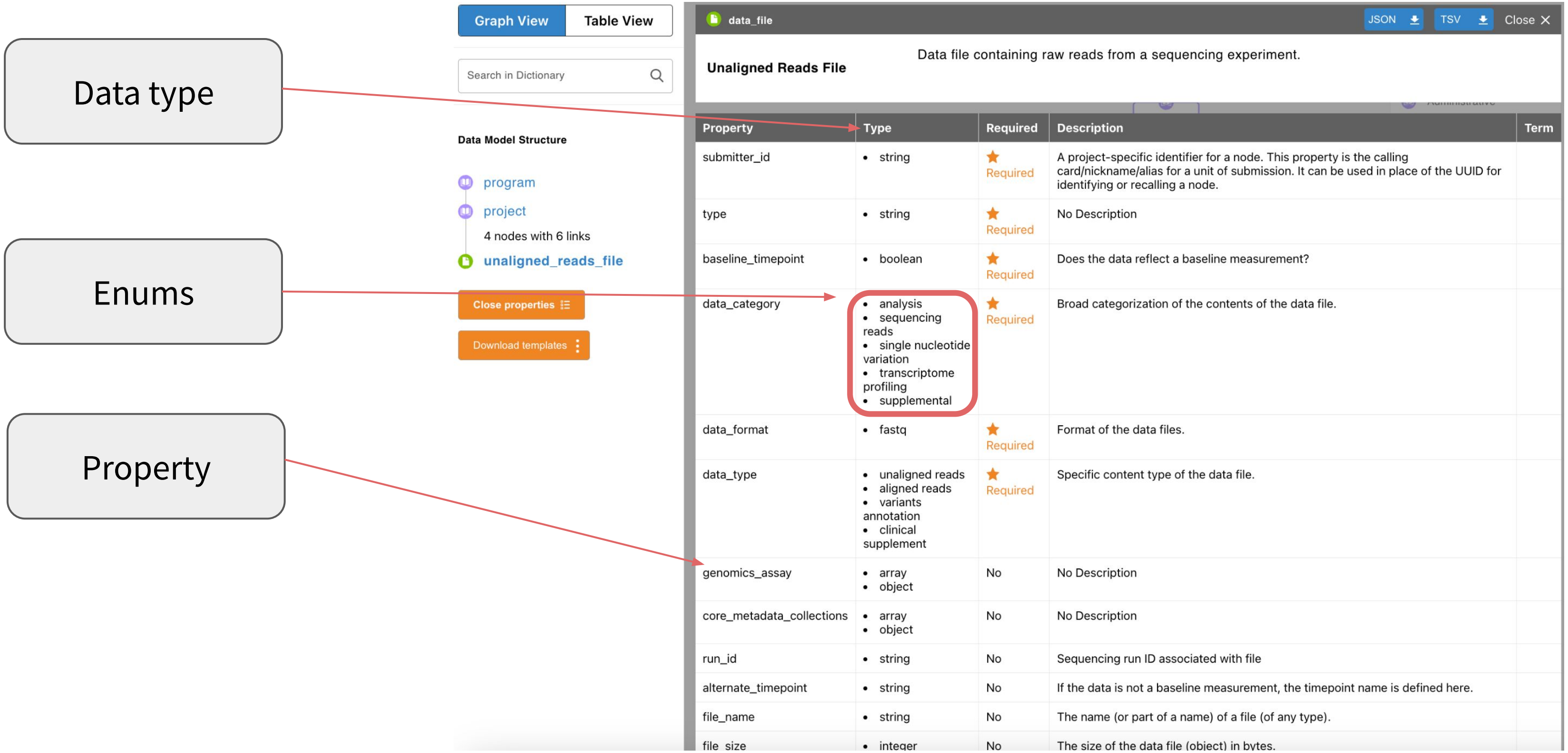
Gen3 data modelling background



Links

Data Object

Gen3 data modelling background



Data Model Structure

- program
- project
- 4 nodes with 6 links
- unaligned_reads_file

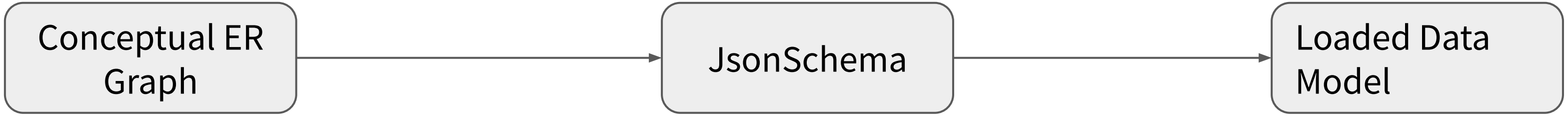
data_file

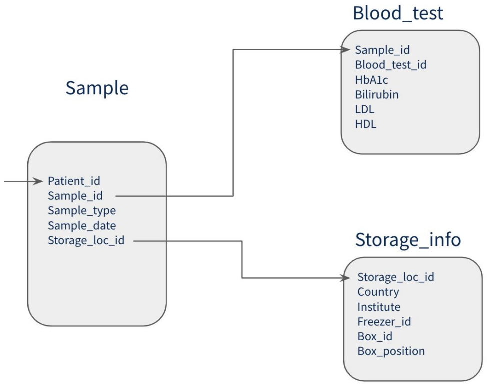
Unaligned Reads File

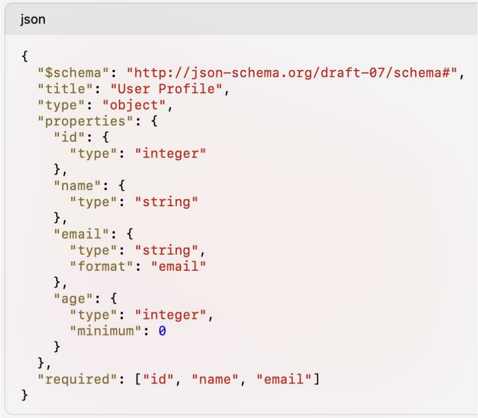
Data file containing raw reads from a sequencing experiment.

Property	Type	Required	Description	Term
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.	
type	• string	★ Required	No Description	
baseline_timepoint	• boolean	★ Required	Does the data reflect a baseline measurement?	
data_category	• analysis • sequencing reads • single nucleotide variation • transcriptome profiling • supplemental	★ Required	Broad categorization of the contents of the data file.	
data_format	• fastq	★ Required	Format of the data files.	
data_type	• unaligned reads • aligned reads • variants annotation • clinical supplement	★ Required	Specific content type of the data file.	
genomics_assay	• array • object	No	No Description	
core_metadata_collections	• array • object	No	No Description	
run_id	• string	No	Sequencing run ID associated with file	
alternate_timepoint	• string	No	If the data is not a baseline measurement, the timepoint name is defined here.	
file_name	• string	No	The name (or part of a name) of a file (of any type).	
file_size	• integer	No	The size of the data file (object) in bytes.	

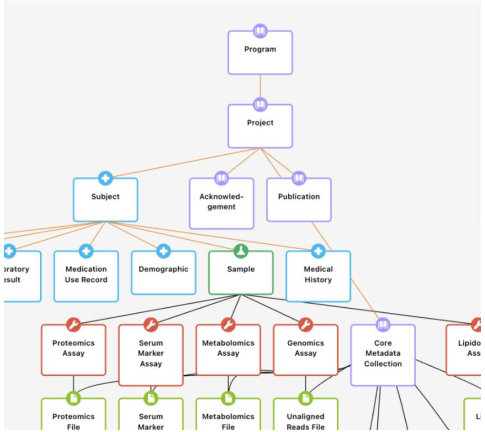
Gen3 data modelling background



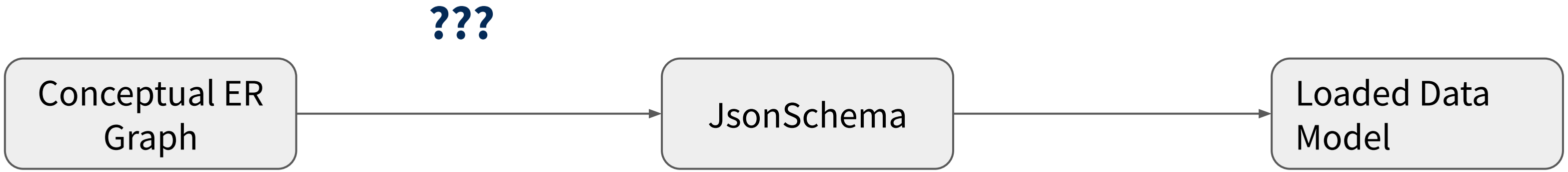




```json
{
  "$schema": "http://json-schema.org/draft-07/schema#",
  "title": "User Profile",
  "type": "object",
  "properties": {
    "id": {
      "type": "integer"
    },
    "name": {
      "type": "string"
    },
    "email": {
      "type": "string",
      "format": "email"
    },
    "age": {
      "type": "integer",
      "minimum": 0
    }
  },
  "required": ["id", "name", "email"]
}
```



Gen3 data modelling background

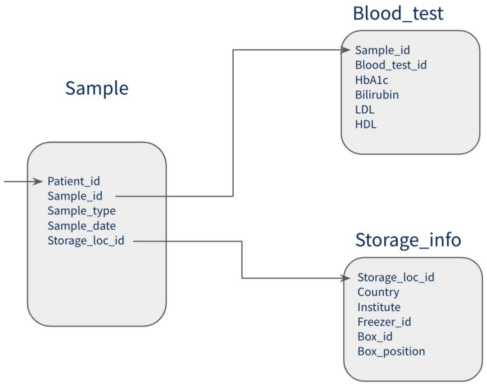


???

Conceptual ER Graph

JsonSchema

Loaded Data Model



Blood_test

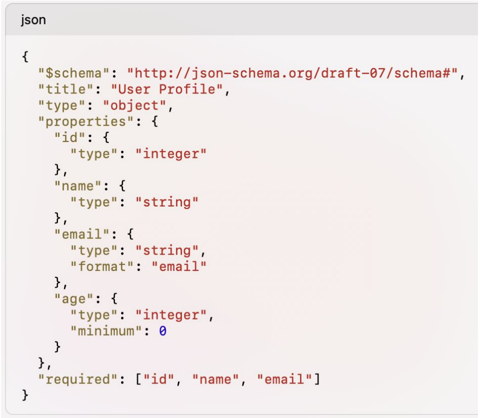
Sample_id
Blood_test_id
HbA1c
Bilirubin
LDL
HDL

Sample

Patient_id
Sample_id
Sample_type
Sample_date
Storage_loc_id

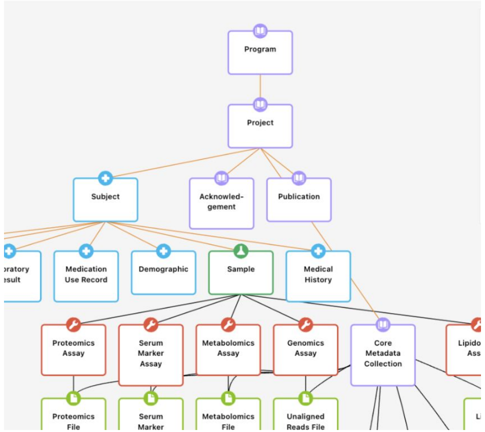
Storage_info

Storage_loc_id
Country
Institute
Freezer_id
Box_id
Box_position



json

```json
{
  "$schema": "http://json-schema.org/draft-07/schema#",
  "title": "User Profile",
  "type": "object",
  "properties": {
    "id": {
      "type": "integer"
    },
    "name": {
      "type": "string"
    },
    "email": {
      "type": "string",
      "format": "email"
    },
    "age": {
      "type": "integer",
      "minimum": 0
    }
  },
  "required": ["id", "name", "email"]
}
```



Thoughts and Experiences with data modelling in Gen3

Thoughts and Experiences with data modelling in Gen3

From a former bioinformatician and beginner data modeller…..

Thoughts and Experiences with data modelling in Gen3

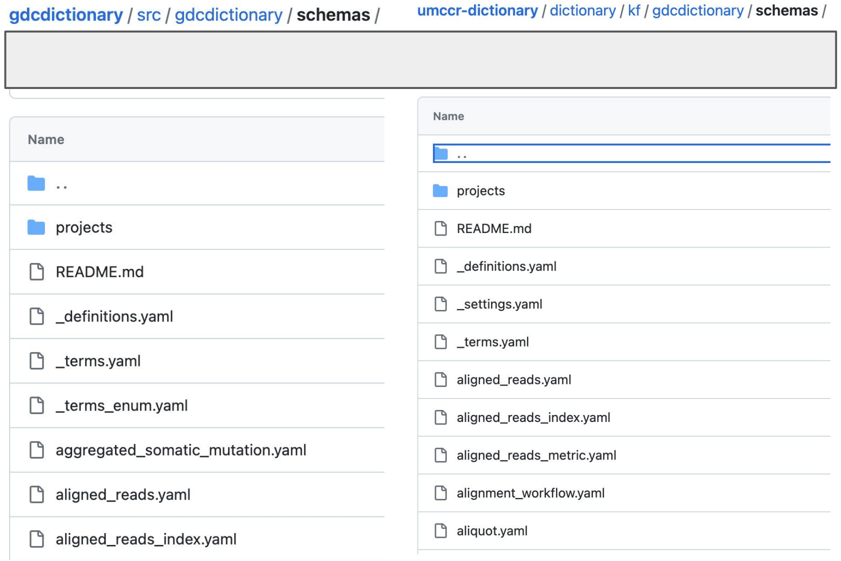
Re-using Data Objects from Other Gen3 Dictionaries

Thoughts and Experiences with data modelling in Gen3

GEN3 DATA COMMONS

Re-using Data Objects from Other Gen3 Dictionaries

- Advantages:
    - Potential for efficiency and consistency
    - Interoperability
    - Many schemas already available in repositories such as
- Challenges:
    - In some cases, this approach leads to dependency and reference complications in the schema



gdcdictionary / src / gdcdictionary / schemas /

Name

- ..
- projects
- README.md
- _definitions.yaml
- _terms.yaml
- _terms_enum.yaml
- aggregated_somatic_mutation.yaml
- aligned_reads.yaml
- aligned_reads_index.yaml

umccr-dictionary / dictionary / kf / gdcdictionary / schemas /

Name

- ..
- projects
- README.md
- _definitions.yaml
- _settings.yaml
- _terms.yaml
- aligned_reads.yaml
- aligned_reads_index.yaml
- aligned_reads_metric.yaml
- alignment_workflow.yaml
- aliquot.yaml

Thoughts and Experiences with data modelling in Gen3

Utilising Common Data Models (CDMs)

Thoughts and Experiences with data modelling in Gen3

Utilising Common Data Models (CDMs)

- Advantages:
    - Adopting ontologies can help other users familiar with that ontology identify groups of data
    - Can promote interoperability with other health data systems
- Limitations
    - Requires a high level of expertise and domain knowledge
    - Utilising a CDM in Gen3 requires conversion tools, e.g. `pfb_fhir` to jsonschema

Thoughts and Experiences with data modelling in Gen3

Building Custom Data Models

Thoughts and Experiences with data modelling in Gen3

Building Custom Data Models

- Efficiency:
    - Building custom data models has provided the fastest and most efficient way of data modelling for our purposes so far
- Pipeline Development:
    - We have set up a data model development pipeline that allows for flexible and frequent updates and testing of our data model prior to acquiring real data

Thoughts and Experiences with data modelling in Gen3

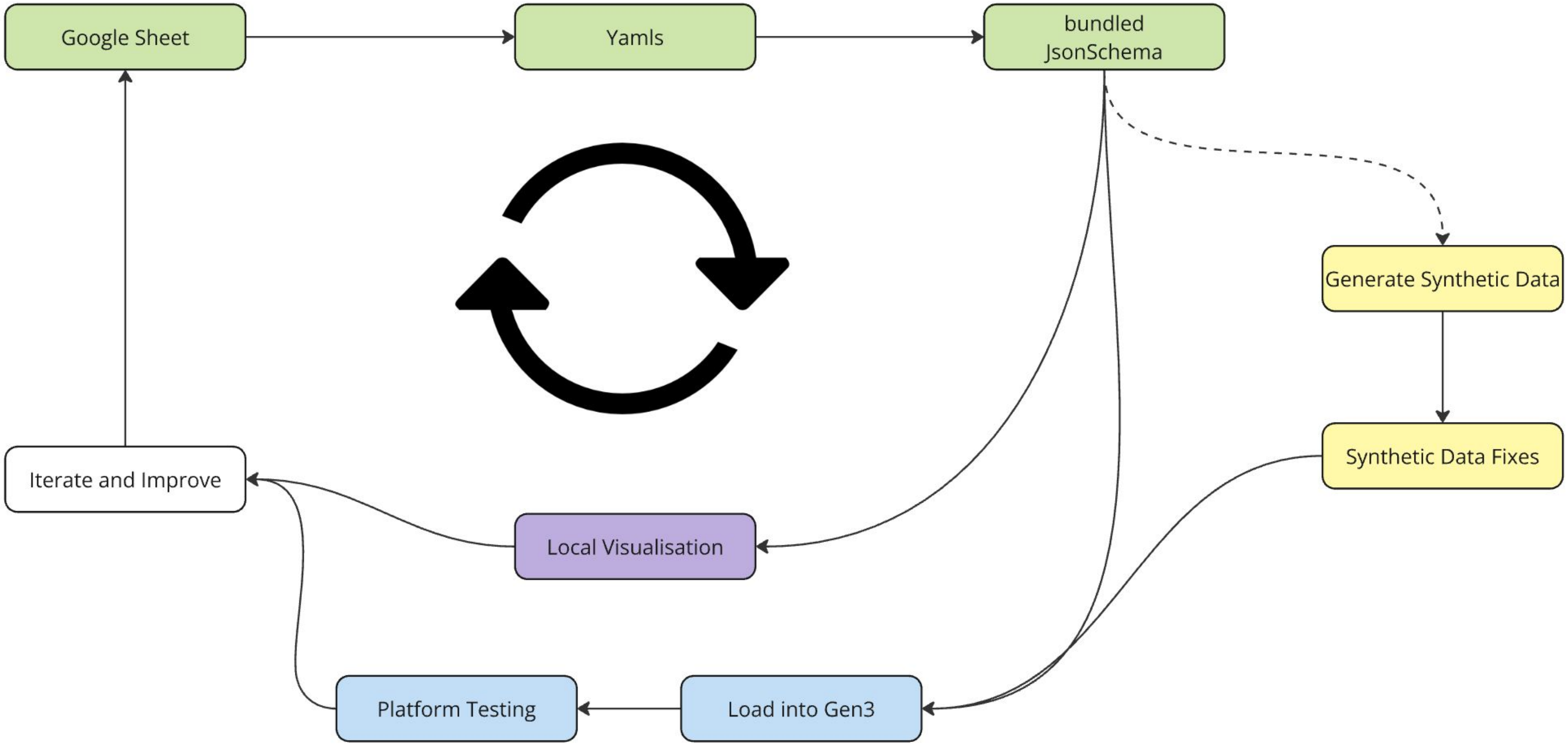
Need for Entry-Level Data Modelling Tools

Thoughts and Experiences with data modelling in Gen3

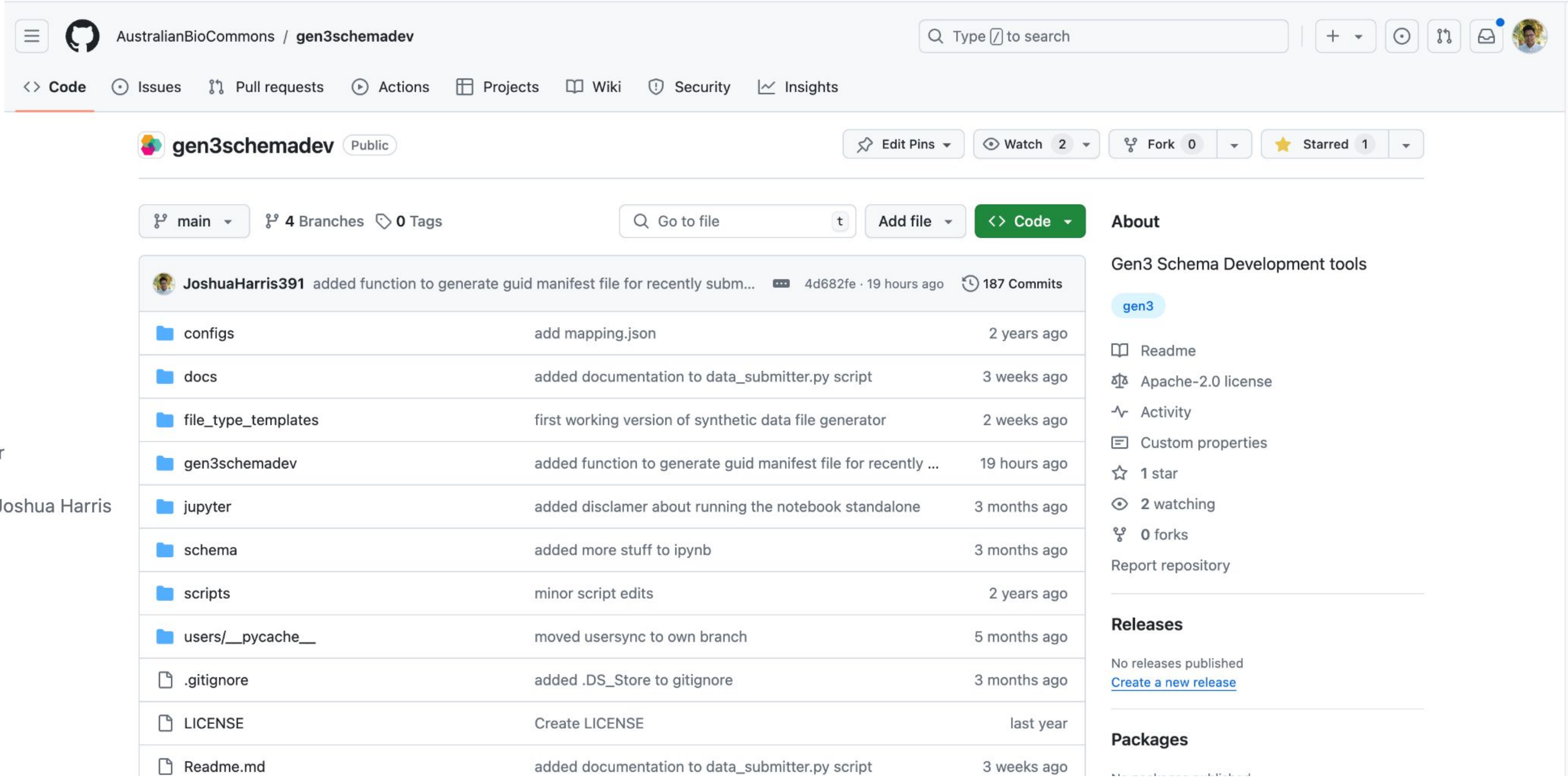
Need for Entry-Level Data Modelling Tools

- Target Audience:
  - Teams of medical researchers or bioinformaticians without extensive experience in data modelling principles and techniques will struggle to adopt gen3
- Challenges with Current Tools:
  - Data modelling with raw JSON schema can be overwhelming for new users
  - We have devised a simplified approach to lower the barrier to entry

Overview of our data modelling process



Australian Biocommons - Gen3schemadev - Git repo



AustralianBioCommons / **gen3schemadev**

gen3schemadev Public

4 Branches · 0 Tags

JoshuaHarris391 added function to generate guid manifest file for recently subm... · 4d682fe · 19 hours ago · 187 Commits

configs	add mapping.json	2 years ago
docs	added documentation to data_submitter.py script	3 weeks ago
file_type_templates	first working version of synthetic data file generator	2 weeks ago
gen3schemadev	added function to generate guid manifest file for recently ...	19 hours ago
jupyter	added disclamer about running the notebook standalone	3 months ago
schema	added more stuff to ipynb	3 months ago
scripts	minor script edits	2 years ago
users/__pycache__	moved usersync to own branch	5 months ago
.gitignore	added .DS_Store to gitignore	3 months ago
LICENSE	Create LICENSE	last year
Readme.md	added documentation to data_submitter.py script	3 weeks ago

About

Gen3 Schema Development tools

gen3

- Readme
- Apache-2.0 license
- Activity
- Custom properties
- 1 star
- 2 watching
- 0 forks

Report repository

Releases

No releases published

Create a new release

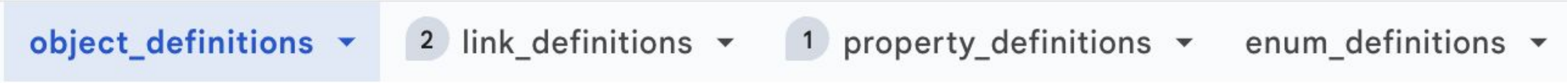
Packages

Contributors 3

- **mshadbolt** Marion
- **uwwint** Uwe Winter
- **JoshuaHarris391** Joshua Harris

Entry Level Data Modelling in Google Sheets

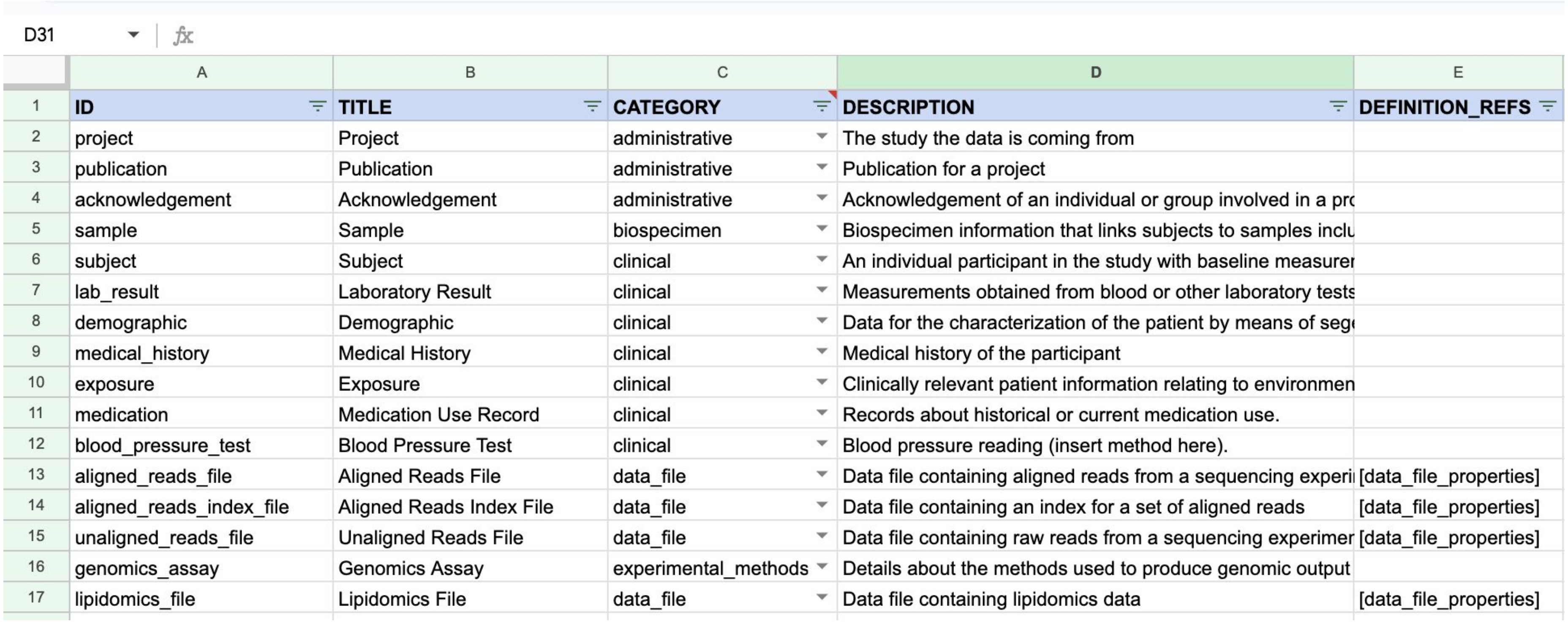
Data Modelling in google sheets utilises 4 main sheets:



Entry Level Data Modelling in Google Sheets

GEN3 DATA COMMONS

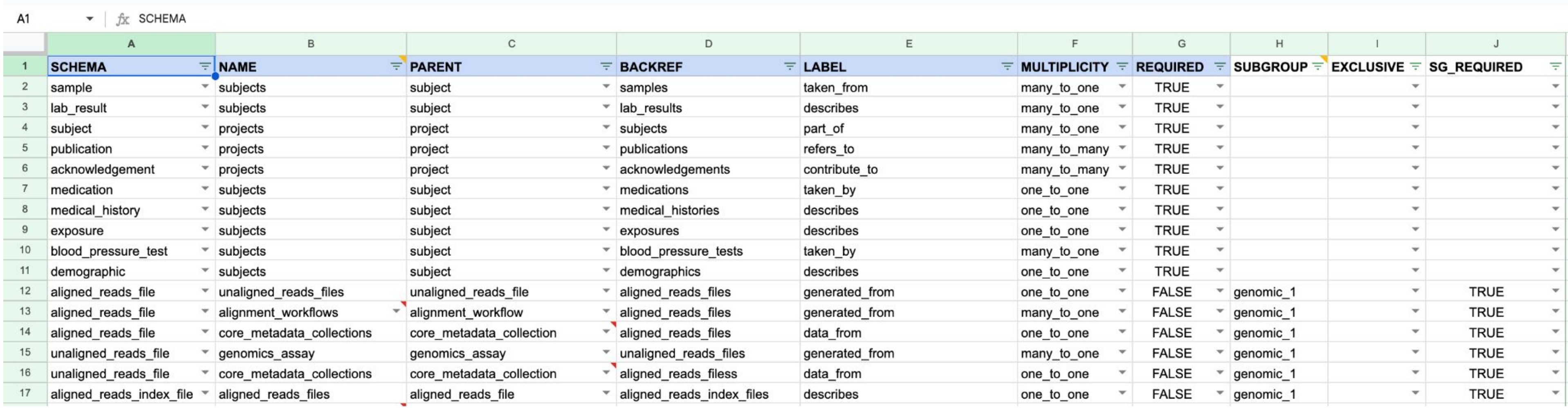
Creating Object Nodes



	A	B	C	D	E
1	ID	TITLE	CATEGORY	DESCRIPTION	DEFINITION_REFS
2	project	Project	administrative	The study the data is coming from	
3	publication	Publication	administrative	Publication for a project	
4	acknowledgement	Acknowledgement	administrative	Acknowledgement of an individual or group involved in a prc	
5	sample	Sample	biospecimen	Biospecimen information that links subjects to samples inclu	
6	subject	Subject	clinical	An individual participant in the study with baseline measurer	
7	lab_result	Laboratory Result	clinical	Measurements obtained from blood or other laboratory tests	
8	demographic	Demographic	clinical	Data for the characterization of the patient by means of seg	
9	medical_history	Medical History	clinical	Medical history of the participant	
10	exposure	Exposure	clinical	Clinically relevant patient information relating to environmen	
11	medication	Medication Use Record	clinical	Records about historical or current medication use.	
12	blood_pressure_test	Blood Pressure Test	clinical	Blood pressure reading (insert method here).	
13	aligned_reads_file	Aligned Reads File	data_file	Data file containing aligned reads from a sequencing experi	[data_file_properties]
14	aligned_reads_index_file	Aligned Reads Index File	data_file	Data file containing an index for a set of aligned reads	[data_file_properties]
15	unaligned_reads_file	Unaligned Reads File	data_file	Data file containing raw reads from a sequencing experimer	[data_file_properties]
16	genomics_assay	Genomics Assay	experimental_methods	Details about the methods used to produce genomic output	
17	lipidomics_file	Lipidomics File	data_file	Data file containing lipidomics data	[data_file_properties]

Entry Level Data Modelling in Google Sheets

Defining Links between Objects

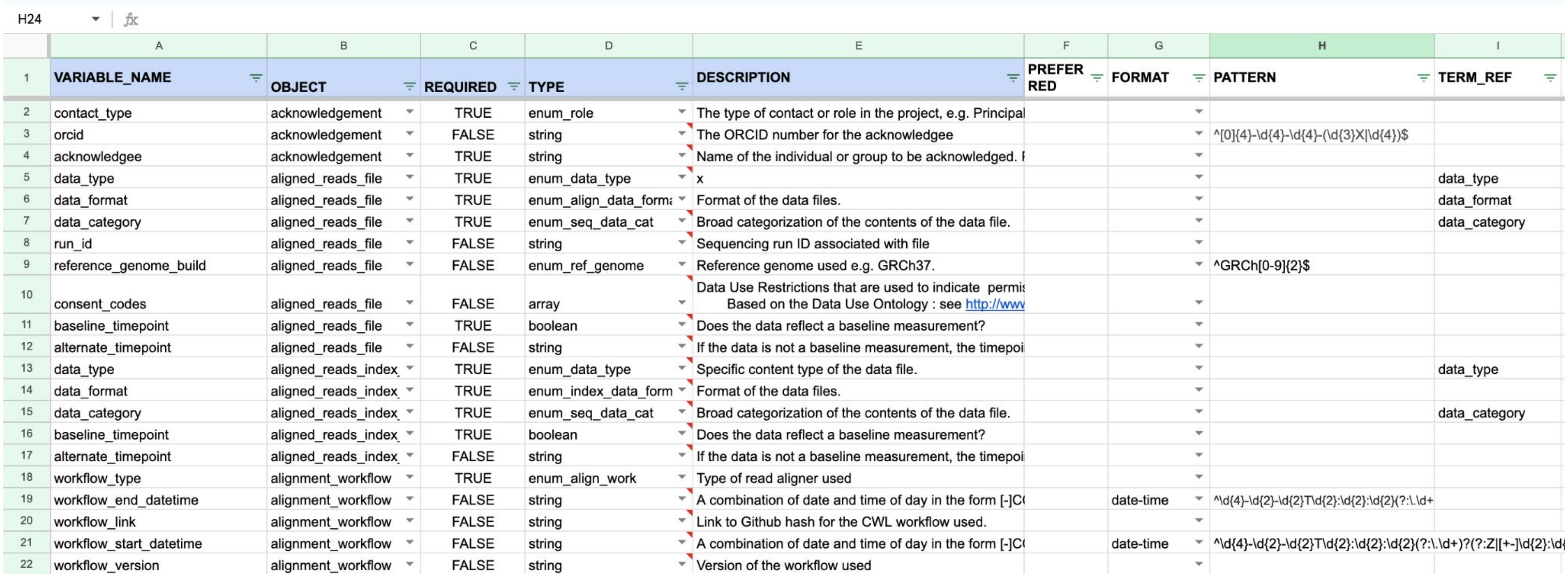


A1 ▾ fx SCHEMA

	A	B	C	D	E	F	G	H	I	J
1	SCHEMA	NAME	PARENT	BACKREF	LABEL	MULTIPLICITY	REQUIRED	SUBGROUP	EXCLUSIVE	SG_REQUIRED
2	sample	subjects	subject	samples	taken_from	many_to_one	TRUE			
3	lab_result	subjects	subject	lab_results	describes	many_to_one	TRUE			
4	subject	projects	project	subjects	part_of	many_to_one	TRUE			
5	publication	projects	project	publications	refers_to	many_to_many	TRUE			
6	acknowledgement	projects	project	acknowledgements	contribute_to	many_to_many	TRUE			
7	medication	subjects	subject	medications	taken_by	one_to_one	TRUE			
8	medical_history	subjects	subject	medical_histories	describes	one_to_one	TRUE			
9	exposure	subjects	subject	exposures	describes	one_to_one	TRUE			
10	blood_pressure_test	subjects	subject	blood_pressure_tests	taken_by	many_to_one	TRUE			
11	demographic	subjects	subject	demographics	describes	one_to_one	TRUE			
12	aligned_reads_file	unaligned_reads_files	unaligned_reads_file	aligned_reads_files	generated_from	one_to_one	FALSE	genomic_1		TRUE
13	aligned_reads_file	alignment_workflows	alignment_workflow	aligned_reads_files	generated_from	many_to_one	FALSE	genomic_1		TRUE
14	aligned_reads_file	core_metadata_collections	core_metadata_collection	aligned_reads_files	data_from	one_to_one	FALSE	genomic_1		TRUE
15	unaligned_reads_file	genomics_assay	genomics_assay	unaligned_reads_files	generated_from	many_to_one	FALSE	genomic_1		TRUE
16	unaligned_reads_file	core_metadata_collections	core_metadata_collection	aligned_reads_filess	data_from	one_to_one	FALSE	genomic_1		TRUE
17	aligned_reads_index_file	aligned_reads_files	aligned_reads_file	aligned_reads_index_files	describes	one_to_one	FALSE	genomic_1		TRUE

Entry Level Data Modelling in Google Sheets

Properties

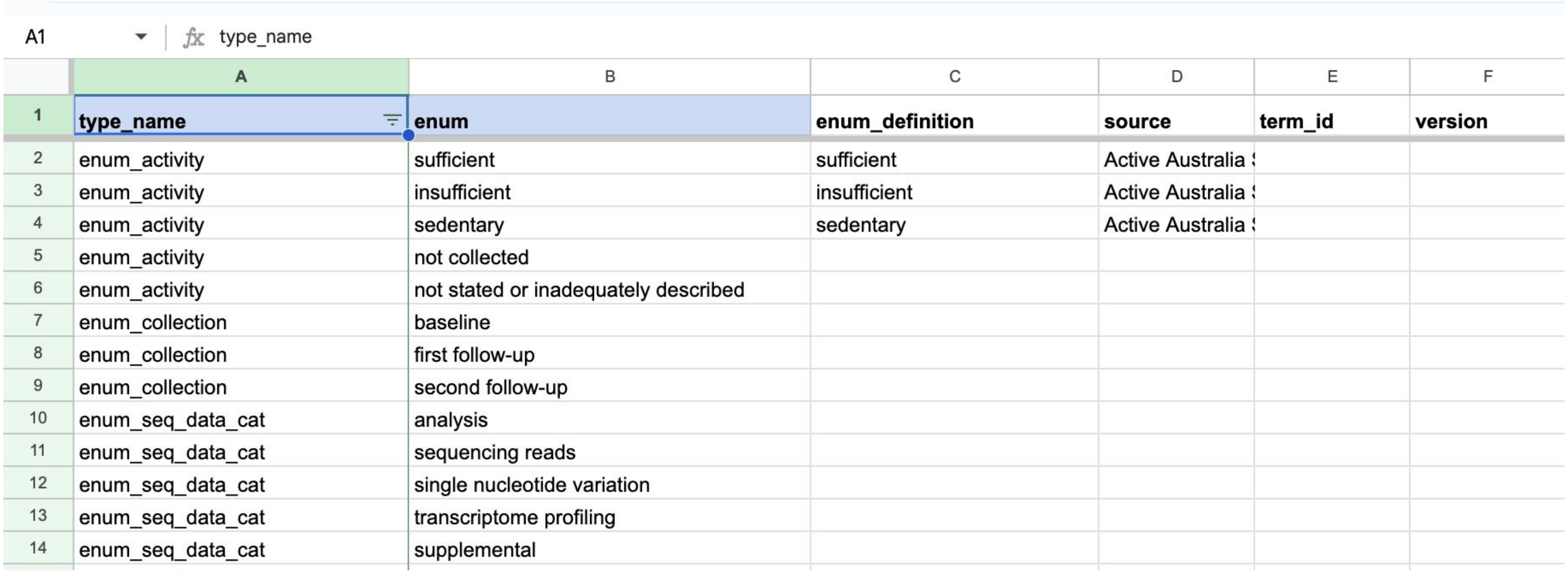


H24 fx

	A	B	C	D	E	F	G	H	I
1	VARIABLE_NAME	OBJECT	REQUIRED	TYPE	DESCRIPTION	PREFERRED	FORMAT	PATTERN	TERM_REF
2	contact_type	acknowledgement	TRUE	enum_role	The type of contact or role in the project, e.g. Principal				
3	orcid	acknowledgement	FALSE	string	The ORCID number for the acknowledgee			^[0]{4}-\d{4}-\d{4}-(\d{3}X|\d{4})$	
4	acknowledgee	acknowledgement	TRUE	string	Name of the individual or group to be acknowledged. F				
5	data_type	aligned_reads_file	TRUE	enum_data_type	x				data_type
6	data_format	aligned_reads_file	TRUE	enum_align_data_forma	Format of the data files.				data_format
7	data_category	aligned_reads_file	TRUE	enum_seq_data_cat	Broad categorization of the contents of the data file.				data_category
8	run_id	aligned_reads_file	FALSE	string	Sequencing run ID associated with file				
9	reference_genome_build	aligned_reads_file	FALSE	enum_ref_genome	Reference genome used e.g. GRCh37.			^GRCh[0-9]{2}$	
10	consent_codes	aligned_reads_file	FALSE	array	Data Use Restrictions that are used to indicate permis Based on the Data Use Ontology : see http://www				
11	baseline_timepoint	aligned_reads_file	TRUE	boolean	Does the data reflect a baseline measurement?				
12	alternate_timepoint	aligned_reads_file	FALSE	string	If the data is not a baseline measurement, the timepoi				
13	data_type	aligned_reads_index_	TRUE	enum_data_type	Specific content type of the data file.				data_type
14	data_format	aligned_reads_index_	TRUE	enum_index_data_form	Format of the data files.				
15	data_category	aligned_reads_index_	TRUE	enum_seq_data_cat	Broad categorization of the contents of the data file.				data_category
16	baseline_timepoint	aligned_reads_index_	TRUE	boolean	Does the data reflect a baseline measurement?				
17	alternate_timepoint	aligned_reads_index_	FALSE	string	If the data is not a baseline measurement, the timepoi				
18	workflow_type	alignment_workflow	TRUE	enum_align_work	Type of read aligner used				
19	workflow_end_datetime	alignment_workflow	FALSE	string	A combination of date and time of day in the form [-]C		date-time	^\d{4}-\d{2}-\d{2}T\d{2}:\d{2}:\d{2}(?:\.\d+	
20	workflow_link	alignment_workflow	FALSE	string	Link to Github hash for the CWL workflow used.				
21	workflow_start_datetime	alignment_workflow	FALSE	string	A combination of date and time of day in the form [-]C		date-time	^\d{4}-\d{2}-\d{2}T\d{2}:\d{2}:\d{2}(?:\.\d+)?(?:Z|[+-]\d{2}:\d	
22	workflow_version	alignment_workflow	FALSE	string	Version of the workflow used				

Entry Level Data Modelling in Google Sheets

Enums

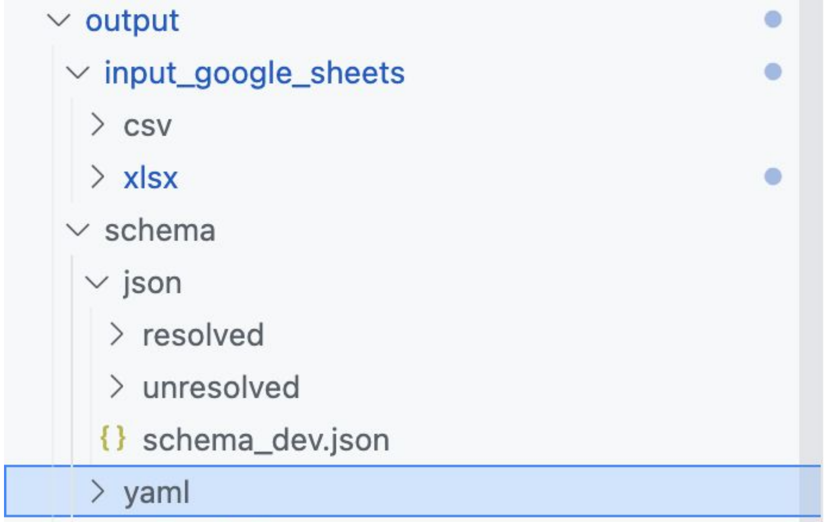


A1 fx type_name

	A	B	C	D	E	F
1	**type_name**	**enum**	**enum_definition**	**source**	**term_id**	**version**
2	enum_activity	sufficient	sufficient	Active Australia		
3	enum_activity	insufficient	insufficient	Active Australia		
4	enum_activity	sedentary	sedentary	Active Australia		
5	enum_activity	not collected				
6	enum_activity	not stated or inadequately described				
7	enum_collection	baseline				
8	enum_collection	first follow-up				
9	enum_collection	second follow-up				
10	enum_seq_data_cat	analysis				
11	enum_seq_data_cat	sequencing reads				
12	enum_seq_data_cat	single nucleotide variation				
13	enum_seq_data_cat	transcriptome profiling				
14	enum_seq_data_cat	supplemental				

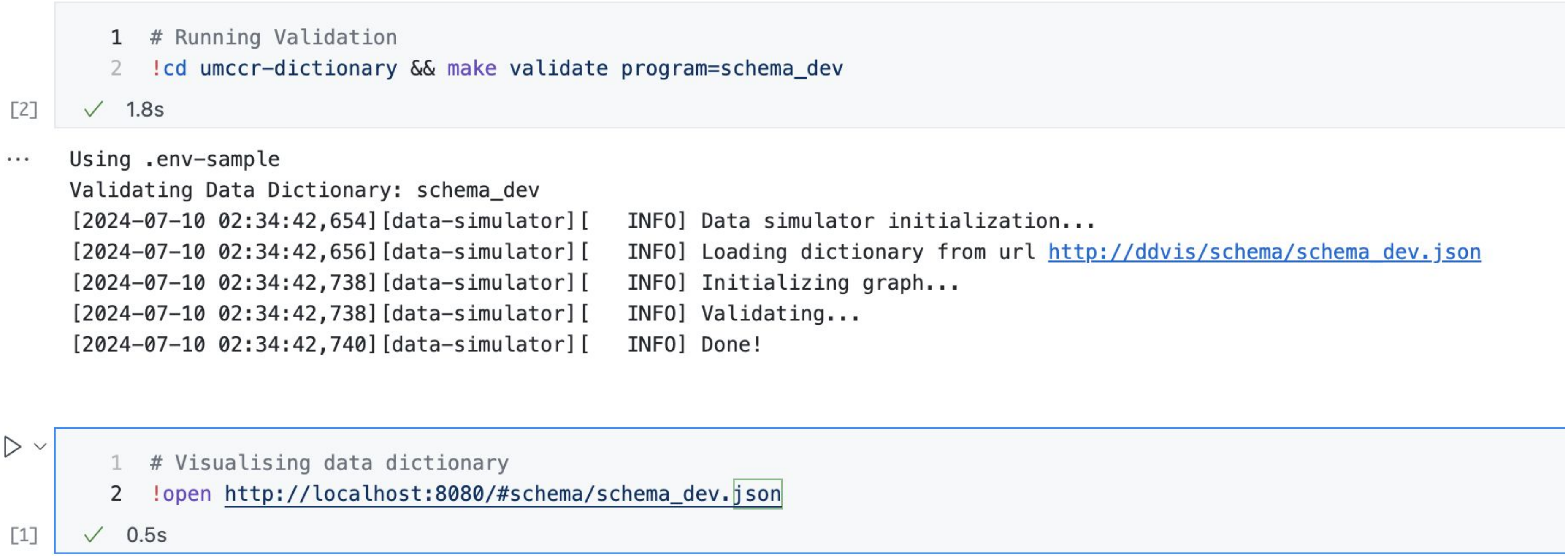
Entry Level Data Modelling in Google Sheets

- `sheet2yaml-CLI.py` Reads google sheets and converts to yamls
- `gen3schemadev` library has functions to also download the current state of the google sheet for your records and reproducibility
- `umccr-g3po` for compiling yamls to jsonschema
-



Validation and Visualisation of Data Model

- Uc-cdis: data-simulator

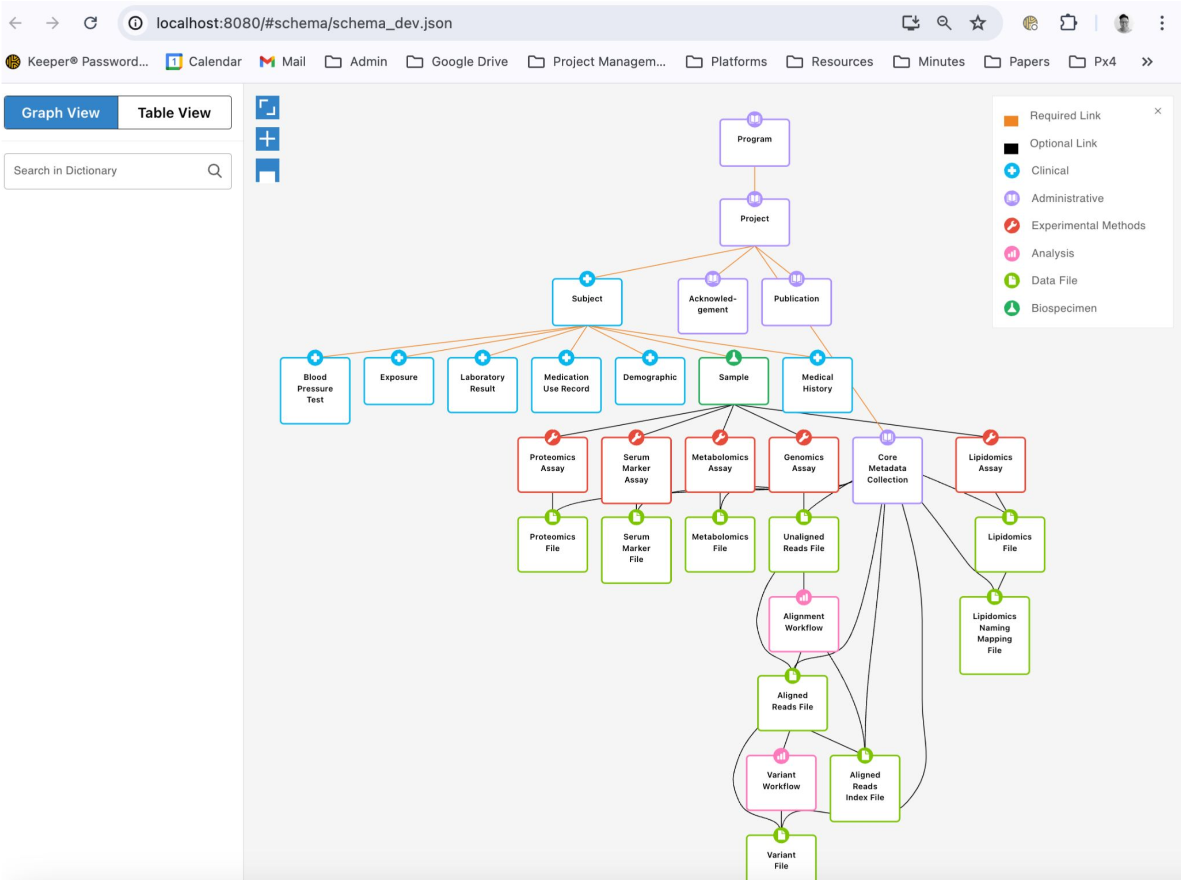


```
# Running Validation
!cd umccr-dictionary && make validate program=schema_dev
```

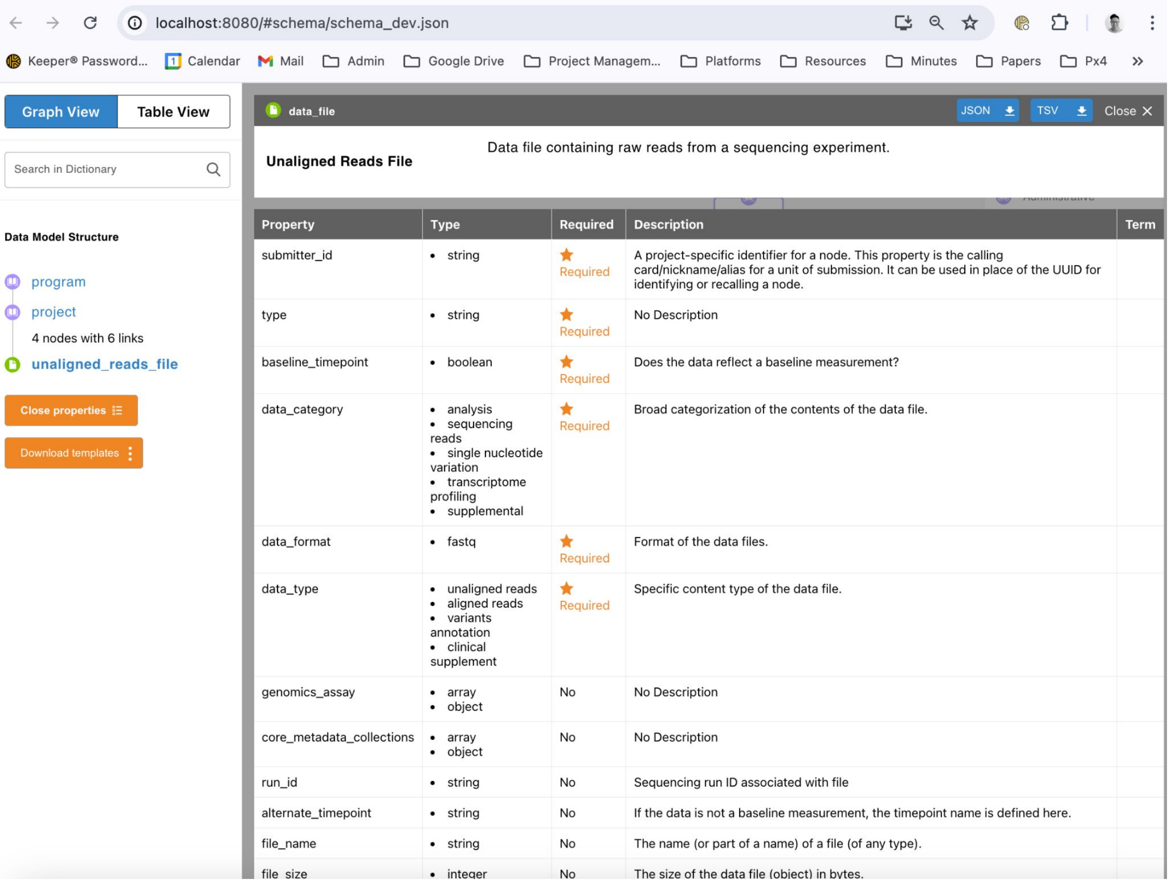
```
Using .env-sample
Validating Data Dictionary: schema_dev
[2024-07-10 02:34:42,654][data-simulator][   INFO] Data simulator initialization...
[2024-07-10 02:34:42,656][data-simulator][   INFO] Loading dictionary from url http://ddvis/schema/schema_dev.json
[2024-07-10 02:34:42,738][data-simulator][   INFO] Initializing graph...
[2024-07-10 02:34:42,738][data-simulator][   INFO] Validating...
[2024-07-10 02:34:42,740][data-simulator][   INFO] Done!
```

```
# Visualising data dictionary
!open http://localhost:8080/#schema/schema_dev.json
```

Validation and Visualisation of Data Model



Validation and Visualisation of Data Model



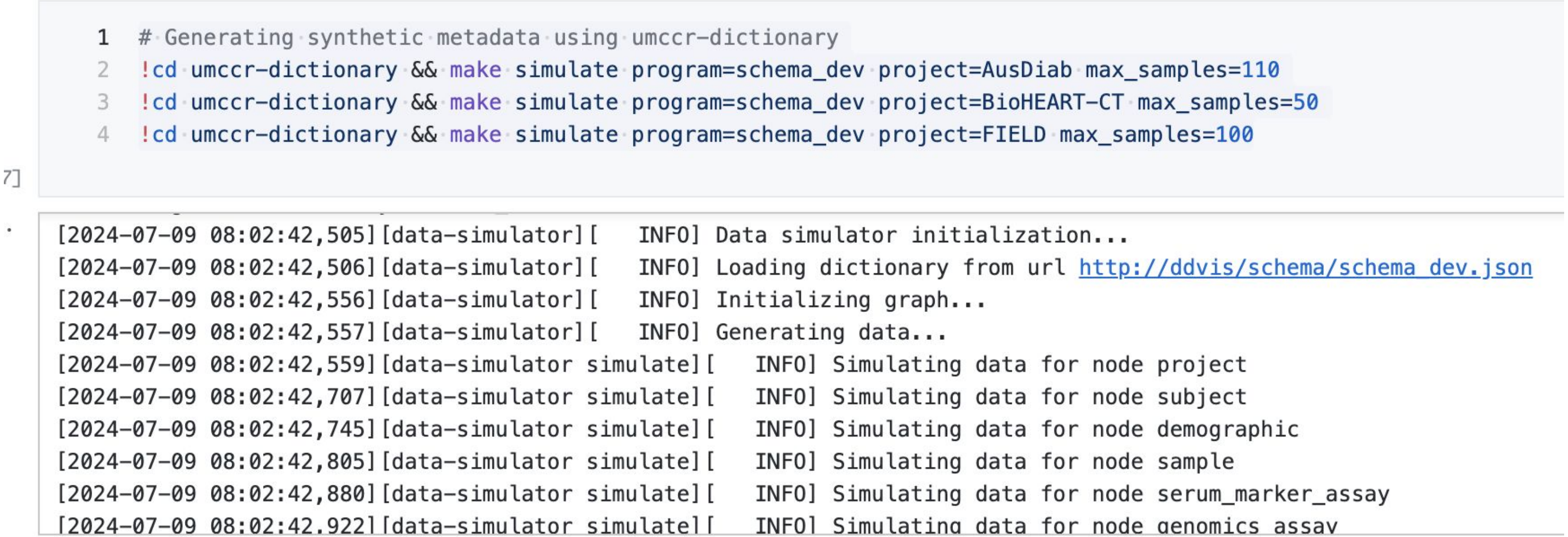
Unaligned Reads File

Data file containing raw reads from a sequencing experiment.

Property	Type	Required	Description	Term
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.	
type	• string	★ Required	No Description	
baseline_timepoint	• boolean	★ Required	Does the data reflect a baseline measurement?	
data_category	• analysis • sequencing reads • single nucleotide variation • transcriptome profiling • supplemental	★ Required	Broad categorization of the contents of the data file.	
data_format	• fastq	★ Required	Format of the data files.	
data_type	• unaligned reads • aligned reads • variants annotation • clinical supplement	★ Required	Specific content type of the data file.	
genomics_assay	• array • object	No	No Description	
core_metadata_collections	• array • object	No	No Description	
run_id	• string	No	Sequencing run ID associated with file	
alternate_timepoint	• string	No	If the data is not a baseline measurement, the timepoint name is defined here.	
file_name	• string	No	The name (or part of a name) of a file (of any type).	
file_size	• integer	No	The size of the data file (object) in bytes.	

Synthetic Data Creation

uc-cdis: data-simulator

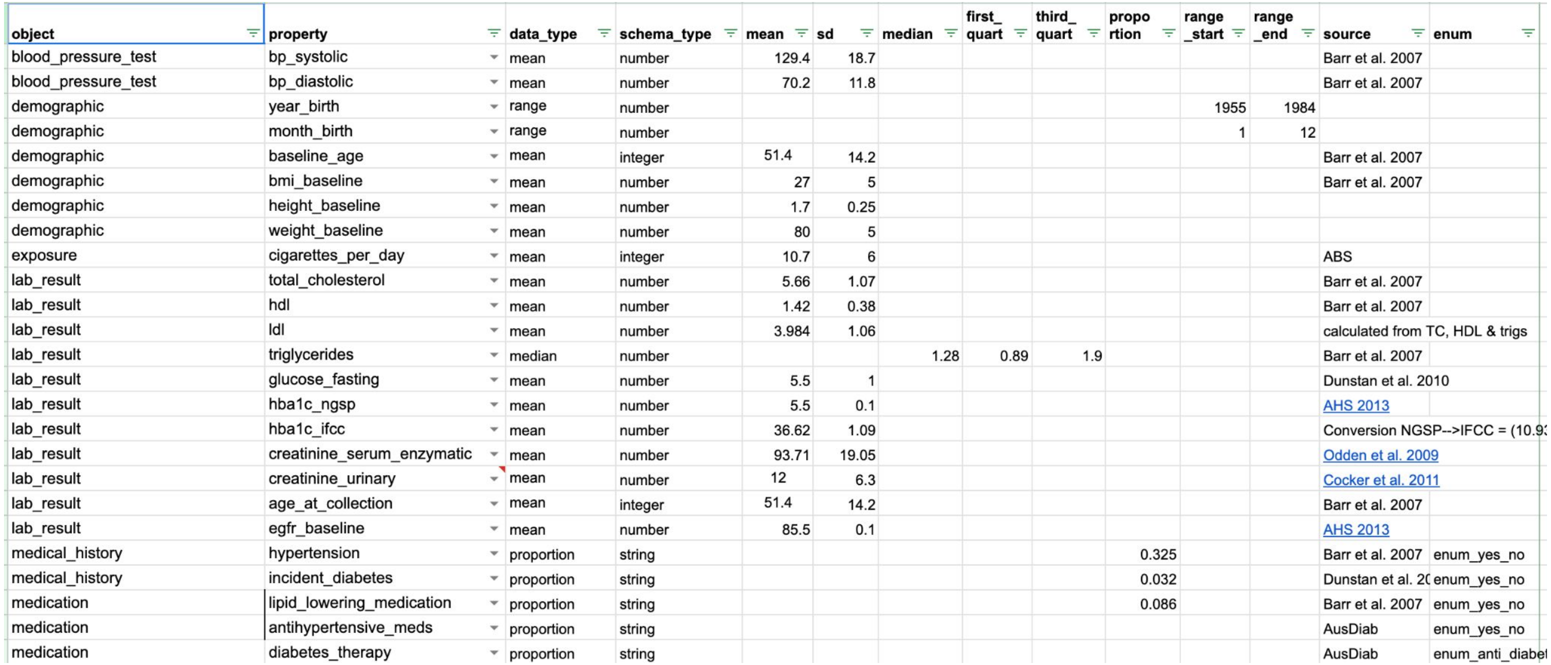


```
# Generating synthetic metadata using umccr-dictionary
!cd umccr-dictionary && make simulate program=schema_dev project=AusDiab max_samples=110
!cd umccr-dictionary && make simulate program=schema_dev project=BioHEART-CT max_samples=50
!cd umccr-dictionary && make simulate program=schema_dev project=FIELD max_samples=100
```

```
[2024-07-09 08:02:42,505][data-simulator][    INFO] Data simulator initialization...
[2024-07-09 08:02:42,506][data-simulator][    INFO] Loading dictionary from url http://ddvis/schema/schema_dev.json
[2024-07-09 08:02:42,556][data-simulator][    INFO] Initializing graph...
[2024-07-09 08:02:42,557][data-simulator][    INFO] Generating data...
[2024-07-09 08:02:42,559][data-simulator simulate][    INFO] Simulating data for node project
[2024-07-09 08:02:42,707][data-simulator simulate][    INFO] Simulating data for node subject
[2024-07-09 08:02:42,745][data-simulator simulate][    INFO] Simulating data for node demographic
[2024-07-09 08:02:42,805][data-simulator simulate][    INFO] Simulating data for node sample
[2024-07-09 08:02:42,880][data-simulator simulate][    INFO] Simulating data for node serum_marker_assay
[2024-07-09 08:02:42,922][data-simulator simulate][    INFO] Simulating data for node genomics_assay
```

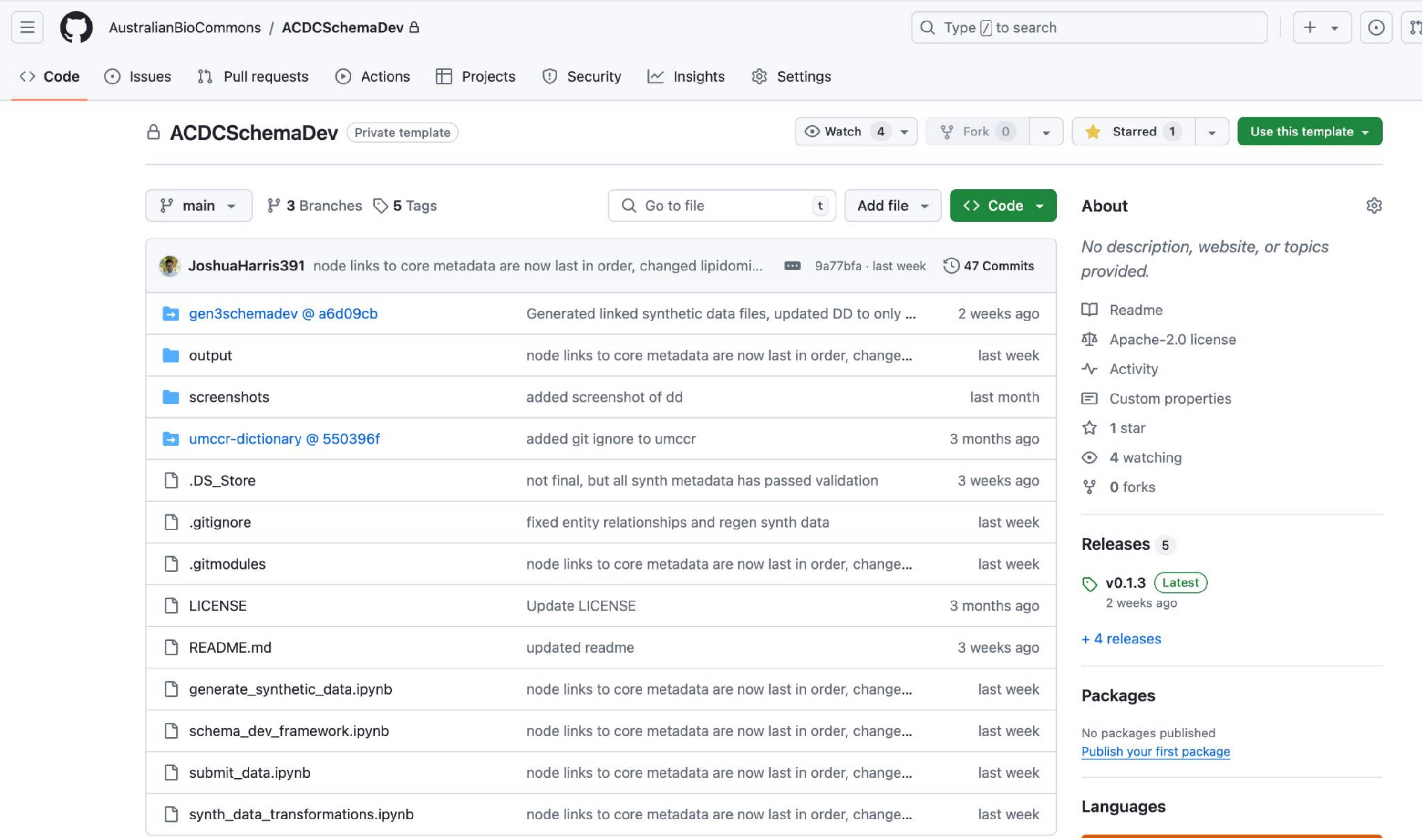
Synthetic Data Creation

Then **gen3schemadev** - plausible_data_gen.py



object	property	data_type	schema_type	mean	sd	median	first_ quart	third_ quart	propo rtion	range _start	range _end	source	enum
blood_pressure_test	bp_systolic	mean	number	129.4	18.7							Barr et al. 2007	
blood_pressure_test	bp_diastolic	mean	number	70.2	11.8							Barr et al. 2007	
demographic	year_birth	range	number							1955	1984		
demographic	month_birth	range	number							1	12		
demographic	baseline_age	mean	integer	51.4	14.2							Barr et al. 2007	
demographic	bmi_baseline	mean	number	27	5							Barr et al. 2007	
demographic	height_baseline	mean	number	1.7	0.25								
demographic	weight_baseline	mean	number	80	5								
exposure	cigarettes_per_day	mean	integer	10.7	6							ABS	
lab_result	total_cholesterol	mean	number	5.66	1.07							Barr et al. 2007	
lab_result	hdl	mean	number	1.42	0.38							Barr et al. 2007	
lab_result	ldl	mean	number	3.984	1.06							calculated from TC, HDL & trigs	
lab_result	triglycerides	median	number			1.28	0.89	1.9				Barr et al. 2007	
lab_result	glucose_fasting	mean	number	5.5	1							Dunstan et al. 2010	
lab_result	hba1c_ngsp	mean	number	5.5	0.1							AHS 2013	
lab_result	hba1c_ifcc	mean	number	36.62	1.09							Conversion NGSP-->IFCC = (10.9	
lab_result	creatinine_serum_enzymatic	mean	number	93.71	19.05							Odden et al. 2009	
lab_result	creatinine_urinary	mean	number	12	6.3							Cocker et al. 2011	
lab_result	age_at_collection	mean	integer	51.4	14.2							Barr et al. 2007	
lab_result	egfr_baseline	mean	number	85.5	0.1							AHS 2013	
medical_history	hypertension	proportion	string						0.325			Barr et al. 2007	enum_yes_no
medical_history	incident_diabetes	proportion	string						0.032			Dunstan et al. 20	enum_yes_no
medication	lipid_lowering_medication	proportion	string						0.086			Barr et al. 2007	enum_yes_no
medication	antihypertensive_meds	proportion	string									AusDiab	enum_yes_no
medication	diabetes_therapy	proportion	string									AusDiab	enum_anti_diabe

Data Model and Synthetic Data Version Management



JoshuaHarris391 node links to core metadata are now last in order, changed lipidomi... 9a77bfa · last week 47 Commits

gen3schemadev @ a6d09cb	Generated linked synthetic data files, updated DD to only ...	2 weeks ago
output	node links to core metadata are now last in order, change...	last week
screenshots	added screenshot of dd	last month
umccr-dictionary @ 550396f	added git ignore to umccr	3 months ago
.DS_Store	not final, but all synth metadata has passed validation	3 weeks ago
.gitignore	fixed entity relationships and regen synth data	last week
.gitmodules	node links to core metadata are now last in order, change...	last week
LICENSE	Update LICENSE	3 months ago
README.md	updated readme	3 weeks ago
generate_synthetic_data.ipynb	node links to core metadata are now last in order, change...	last week
schema_dev_framework.ipynb	node links to core metadata are now last in order, change...	last week
submit_data.ipynb	node links to core metadata are now last in order, change...	last week
synth_data_transformations.ipynb	node links to core metadata are now last in order, change...	last week

About

No description, website, or topics provided.

- Readme
- Apache-2.0 license
- Activity
- Custom properties
- 1 star
- 4 watching
- 0 forks

Releases 5

v0.1.3 Latest
2 weeks ago

+ 4 releases

Packages

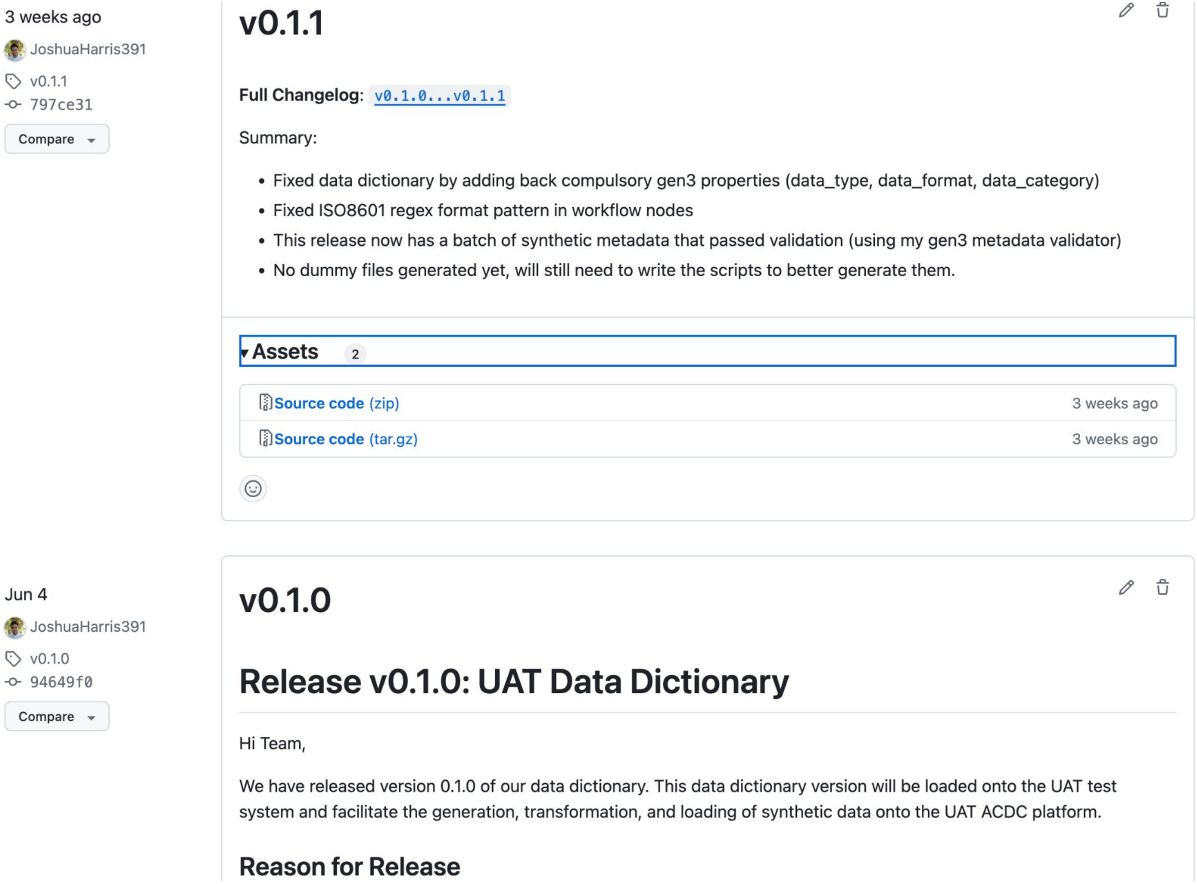
No packages published
Publish your first package

Languages

Data Model and Synthetic Data Version Management

GEN3 DATA COMMONS

We manage our schema version and matching synthetic data batches with git releases



3 weeks ago

JoshuaHarris391

v0.1.1

797ce31

Compare

v0.1.1

Full Changelog: `v0.1.0...v0.1.1`

Summary:

- Fixed data dictionary by adding back compulsory gen3 properties (data_type, data_format, data_category)
- Fixed ISO8601 regex format pattern in workflow nodes
- This release now has a batch of synthetic metadata that passed validation (using my gen3 metadata validator)
- No dummy files generated yet, will still need to write the scripts to better generate them.

▾ Assets 2

Source code (zip)	3 weeks ago
Source code (tar.gz)	3 weeks ago

Jun 4

JoshuaHarris391

v0.1.0

94649f0

Compare

v0.1.0

Release v0.1.0: UAT Data Dictionary

Hi Team,

We have released version 0.1.0 of our data dictionary. This data dictionary version will be loaded onto the UAT test system and facilitate the generation, transformation, and loading of synthetic data onto the UAT ACDC platform.

Reason for Release

Benefits and Limitations of our method

- Advantages:
    - Easy to use
    - Low barrier of entry
    - Good for prototyping
    - Can help you create the bulk of your data model before working explicitly with jsonschema
    - google sheets used to compile the json schema is saved for reproducibility
- Limitations
    - Not reverse compatible (json schema -> google sheet)
    - Can only incorporate CDM elements or other gen3 data objects after compilation to jsonschema

Future Work

- Potentially package this workflow and tools into an open source project
- Reverse engineer jsonschema back to google sheet
- Finalise tools for gen3 data model node/object ingestion

Acknowledgements

Funders:

- BPA - Bioplatforms Australia

BIOPLATFORMS
AUSTRALIA

- MRFF - Medical Research Future Fund - Australia

Medical Research
Future Fund

Acknowledgements

- Speakers
  - Nathalie Giraudon, New Zealand eScience Infrastructure (NeSI)
  - Plamen Martinov, Open Commons Consortium
  - Chris Meyer, Center for Translational Data Science, University of Chicago
  - Liam Beckman, Oregon Health and Science University
  - Joshua Harris, Australian BioCommons
- 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