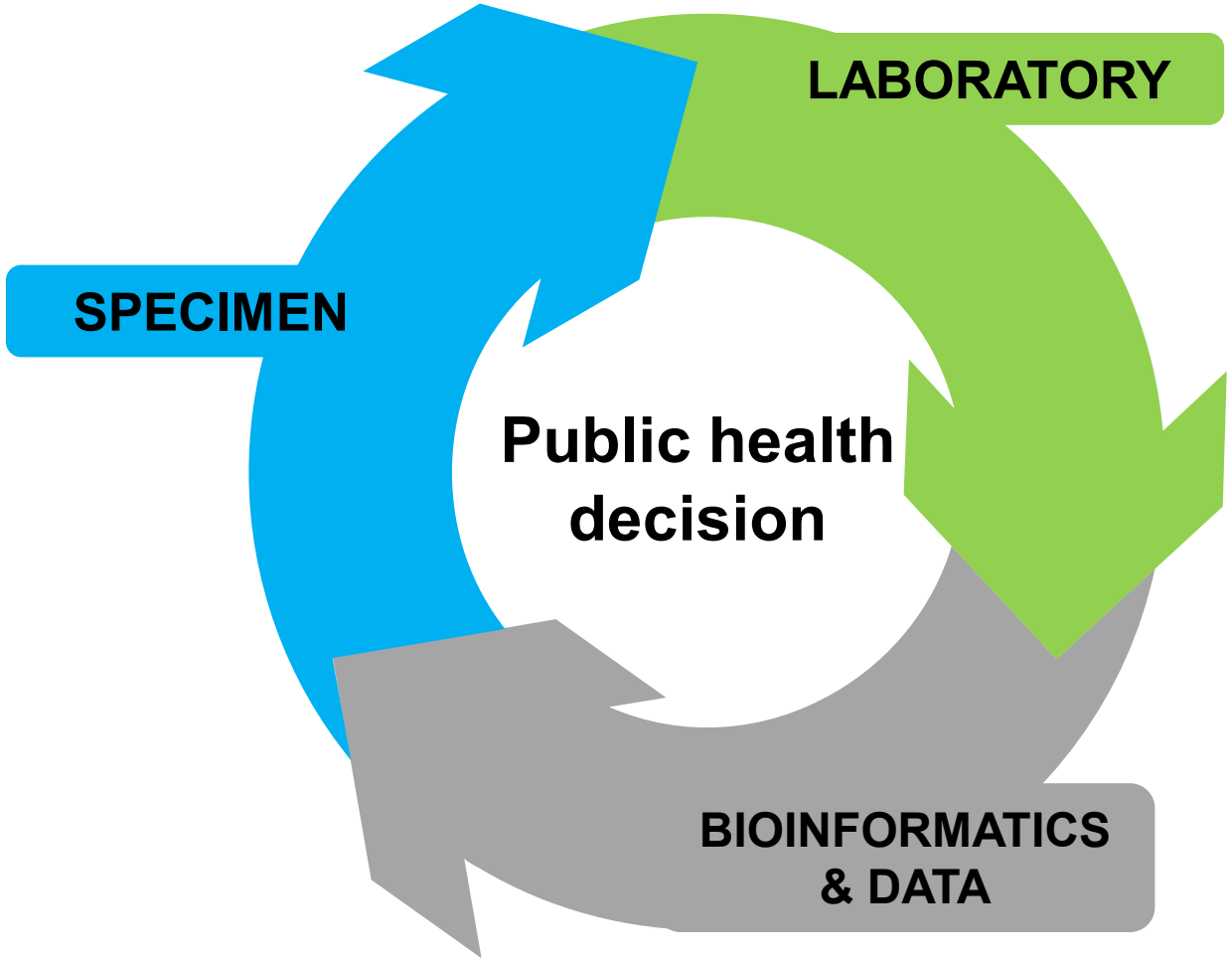
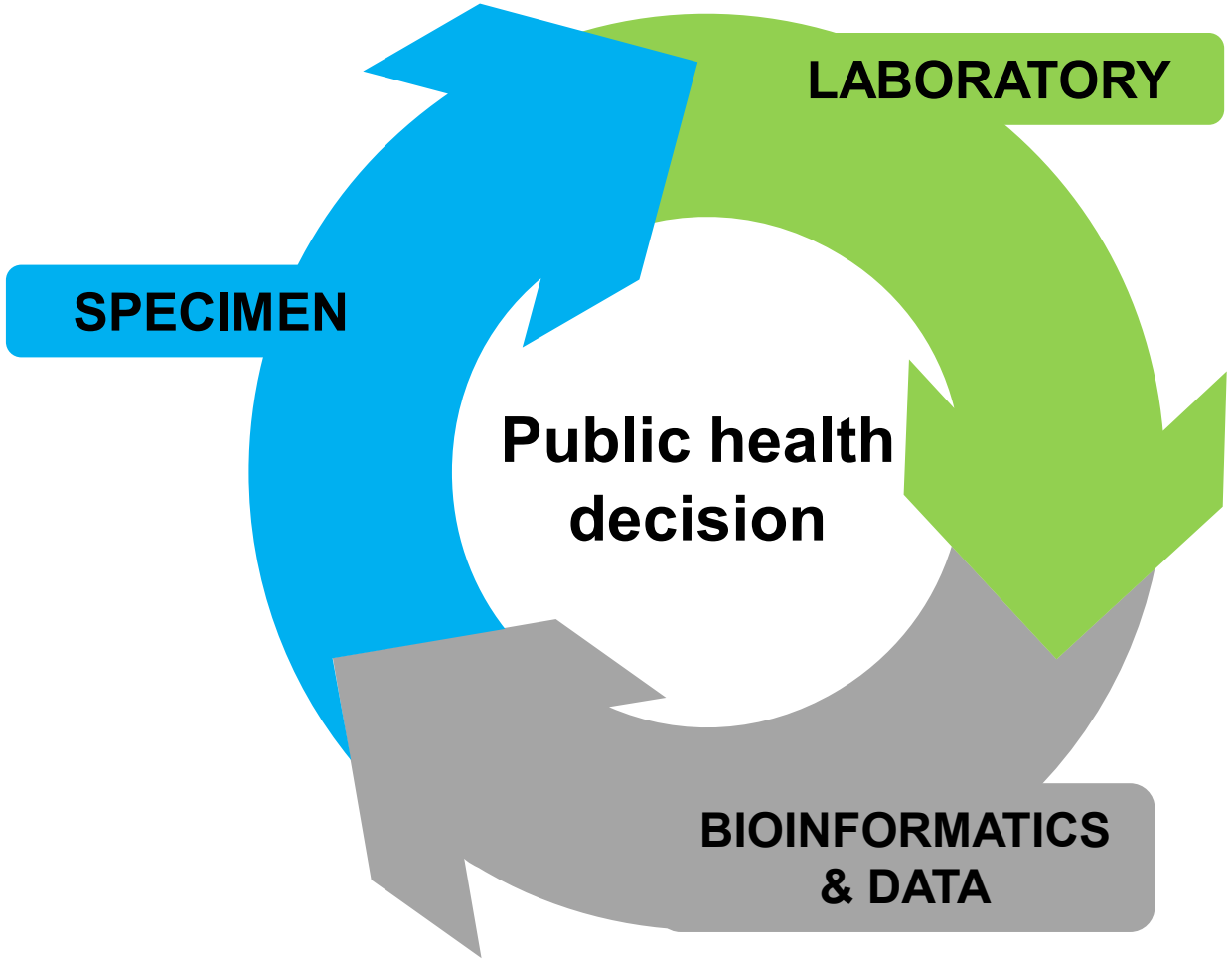
Data organization & meta data capture

Alan Christoffels

Data organisation



Data organisation



- Experimental data/measurements
- Sequencing prep
- Sample info
- Files
- workflows

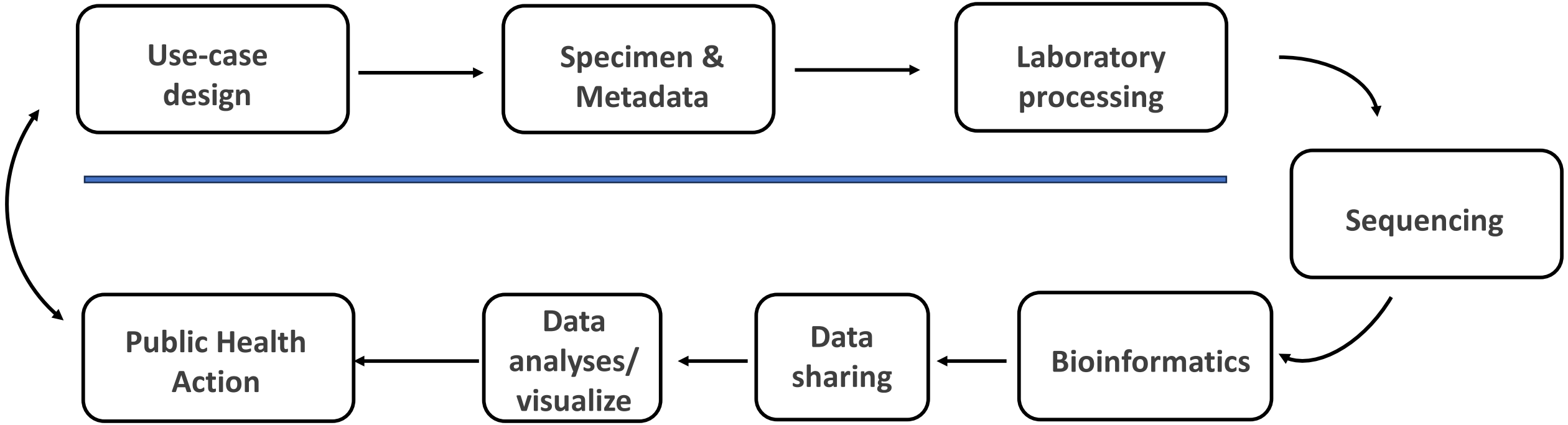
Objectives for this session

- How do we structure meta data
- Difference between Ontology and data standards
- PHA4GE SARS-COV-2 meta data standard specification

Discussion #1

- What kinds of data and information have you generated before you sent your DNA/RNA off for sequencing?

Laboratory Workflows



Data that we capture before sequencing

- Spreadsheet with data from an experiment
- Lab notebook with experimental info
- Speadsheet about the samples you sent off
- Notes about sequencing prep
- Type of sequencing
- Importance of dates when stuff was done → helps for unique identifiers
  - Tracking of samples

Capture meta data in Spreadsheets

- Leave the raw data raw - do not change it!
- Put each observation or sample in its own row.
- Put all your variables in columns - the thing that vary between samples, like ‘strain’ or ‘DNA-concentration’.
- column names must be explanatory and no space eg., LibraryPrep or Library_Prep
- Do not combine multiple pieces of information in once cell
  - E.coli_K12 → rather have a column for species (E.coli) and a strain column (K12)
- Export the data to a text-based format like comma-separated values

Discussion #2: work in pairs: What's wrong with this?

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	
1	***REL606A: generation 0***														
2	**BLOUNT et al. 20**	sequenced at MSU	single end, 35 or 36 bp						**Tenaillon et al. 2(**	Population: Ara-3					
3	STRAIN	Generation	Clade	Population	Mtator	Run	Sequencing depth	CIT	strain	generation	clade	mutator	facility	run	read
4	ZDB464	20000	(C1,C2)	ara-3	None	SRR098285	29,7	unknown	REL2181A	5000		none	MSU RTSF	SRR2589044	pair
5	REL10979	40000	C3+H	ARA-3	"+"	SRR098029	30,1	"+"	REL966A	1000		none	IntraGen	SRR2589001	pair
6	REL10988	40000	C2	Ara-3	"+"	SRR098030	30,2	minus	REL764B	500		None	IntraGen	SRR2584853	pair
7	ZDB429	10000	UC	Ara-3	None	SRR098282	87,3	unknown	REL1166A	2000		None	IntraGen	SRR2584859	pair
8	ZDB357	30000	C2	Ara-3	None	SRR098280	111,2	unknown	REL7179B	15000		None	MSU RTSF	SRR2584863	pair
9	ZDB16	30000	C1	Ara-3	None	SRR098031	113,9	unknown	REL1070A	1500		None	IntraGen	SRR2584857	pair
10	ZDB458	20000	(C1,C2)	Ara-3	None	SRR098284	126,8	unknown	REL4538A	10000	UC	None	MSU RTSF	SRR2589045	pair
11	ZDB446	15000	UC	Ara-3	None	SRR098283	141,1	unknown	REL966B	1000		None	IntraGen	SRR2584856	pair
12	ZDB409	5000	unknown	Ara-3		SRR098281	144,2	unknown	REL1070B	1500		None	IntraGen	SRR2584858	
13	ZDB467	20000	(C1,C2)	Ara-3		SRR098286		unknown	REL1166B	2000		None	MSU RTSF	SRR2591041	
14	ZDB477	25000	C1	Ara-3		SRR098287		unknown	REL764A	500		None	IntraGen	SRR2584852	pair
15	ZDB483	25000	C3	Ara-3		SRR098288		unknown	REL11365	50000	C3+H	plus	MSU RTSF	SRR2584866	pair
16	ZDB199	31500	C1	Ara-3	None	SRR098044		minus	REL11364	50000	C3+H	plus	MSU RTSF	SRR2584864	sing
17	ZDB200	31500	C2	Ara-3	None	SRR098279		minus							
18	ZDB564	31500	C3+	Ara-3	None	SRR098289		"+"	**Leon et al. 2018**	read length: 101;	population: ARA-3	clade: C3	sequenced at UTA GSAF		
19	ZDB172	32000	C3+	Ara-3	None	SRR098042		plus	**strain**	**generation**	**run**	**cit**			
20	ZDB30	32000	C3	Ara-3	None	SRR098032		minus	ZDB1	10000	SRR6178299	unk			
21	ZDB143	32500	C2	Ara-3	None	SRR098041		minus	ZDB425	10000	SRR6178304	unk			
22	ZDB158	32500	C2	Ara-3	None	SRR098040		minus	ZDB445	15000	SRR6178301	unk			
23	CZB152	33000	C3+	Ara-3	None	SRR098027		lus	ZDB478	25000	SRR6178302	unk			
24	CZB154	33000	C3+	Ara-3	None	SRR097977		plus	ZDB486	25000	SRR6178309	unk			
25	CZB199	33000	C1	Ara-3	None	SRR098026		minus	ZDB488	25000	SRR6178310	unk			
26	ZDB83	34000	C3+	Ara-3	None	SRR098034		plus	ZDB309	27000	SRR6178307	unk			
27	ZDB87	34000	C2	Ara-3	None	SRR098035		minus	ZDB310	27000	SRR6178308	unk			
28	ZDB96	36000	C3+H	Ara-3	plus	SRR098036		plus	ZDB317	27000	SRR6178305	unk			
29	ZDB99	36000	C1	Ara-3	None	SRR098037		minus	ZDB334	28000	SRR6178306	unk			
30	ZDB107	38000	C3+H	Ara-3	plus	SRR098038		plus	ZDB339	28000	SRR6178303	unk			
31	ZDB111	38000	C2	Ara-3	None	SRR098039		minus	ZDB13	29000	SRR6178300	unk			
32									ZDB14	29000	SRR6178297	unk			
33									ZDB17	30000	SRR6178298	unk			

Work in pairs: Data capture for sequencing: what is wrong with this data?

well_position	tube_barcode	plate_barcode	client_sample_id		replicate	Volume (¬µL)	concentration (ng/¬µL)	RIN	prep_date	ship_date
A1	151017990	LP-10624	wild type 1h1	a	64.2	211.07	8.1		6-Jul-15	20-Jul
B1	151101577	LP-10624	wild type 1h1	b	63.7	220.21	9.4		6-Jul-15	20-Jul
C1	151142725	LP-10624	wild type 1h1	c	60.2	207.57	8.9		6-Jul-15	20-Jul
D1	151232891	LP-10624	wild type 1h-2	A	55.8	180.62	9		6-Jul-15	20-Jul
E1	151236606	LP-10624	wild type 1h-2	B	60.8	190.86	8.1		6-Jul-15	20-Jul
F1	151323716	LP-10624	wild type 1h-2	C	57.5	192.97	8.6		6-Jul-15	20-Jul
G1	151346588	LP-10624	wild type 1h-3	A	64.9	218.88	8.6		6-Jul-15	20-Jul
H1	151423653	LP-10624	wild type 1h-3	B	62.5	173.44	8.8		6-Jul-15	20-Jul
A2	151462684	LP-10624	wild type 1h-3	C	53.9	214.11	9.5		6-Jul-15	20-Jul
B2	151508377	LP-10624	wild type 1h-4	A	62.4	209.63	8.1		6-Jul-15	20-Jul
C2	151539039	LP-10624	wild type 1h-4	B	66	222.44	8.8		6-Jul-15	20-Jul
D2	151545962	LP-10624	wild type 1h-4	C	61.5	206.27	8		6-Jul-15	20-Jul
E2	151588038	LP-10624	wild type 1h-5	A	58.2	157.67	8.9		6-Jul-15	20-Jul
F2	151666965	LP-10624	wild type 1h-5	B	68	206.45	8.3		6-Jul-15	20-Jul
G2	151719126	LP-10624	wild type 1h-5	C	56.6	220.84	8.4		6-Jul-15	20-Jul
H2	151767622	LP-10624	wild type 1h-6	A	54	179.47	8.3		6-Jul-15	20-Jul
A3	151781088	LP-10624	wild type 1h-6	B	59.6	197.08	8.5		6-Jul-15	20-Jul
B3	151796026	LP-10624	wild type 1h-6	C	56.8	219.34	8		6-Jul-15	20-Jul
C3	151882778	LP-10624	wild type 1h-7	A	57.2	182.17	7.9		7-Jun-15	20-Jul
D3	151944346	LP-10624	wildtype 1h-7	B	630.1	186.98	9.2		7-Jun-15	20-Jul
E3	151970881	LP-10624	wildtype 1h-7	C	63.4	194.28	8.5		7-Jun-15	20-Jul

Data capture for sequencing: what is wrong?

- Format of client_sample_id changes and cannot have spaces
- Capitalization of the replicate column changes
- Volume and concentration column headers have unusual (not allowed) characters
- Volume, concentration, and RIN column decimal accuracy changes
- The prep_date and ship_date formats are different, and prep_date has multiple formats
- Missing data

Challenges of Heterogenous data

- Field names could be be used differently
    - Source (lab) versus source (sample type)
- Values: usually free text
    - Short hand
    - Granularity (cough vs dry cough)
    - Format of date

Getting the right information to the right people is critical during health emergencies.

- Data structure variability in local databases propagates to public repositories

Private databases:

Specimen Collected
☐ Upper respiratory (e.g., Nasopharyngeal or oropharyngeal swab)
☐ Lower respiratory (e.g., sputum, tracheal aspirate, BAL, pleural fluid)

6 - Specimen Type (check all that apply)

Specimen Collection Date: yyyy / mm / dd **(required)**

☐ NPS in UTM

☐ Throat Swab in UTM

☐ Other (Specify):

If possible:

☐ BAL

☐ Sputum

Public databases:

isolate	SARS-CoV-2/186197/human/2020/Malaysia
collected by	Universiti Malaya COVID Research group
collection date	14-Mar-2020
geographic location	Malaysia
host	Homo sapiens
host disease	COVID-19
isolation source	Nasopharyngeal/throat swab
latitude and longitude	3.1390 N 101.6869 E

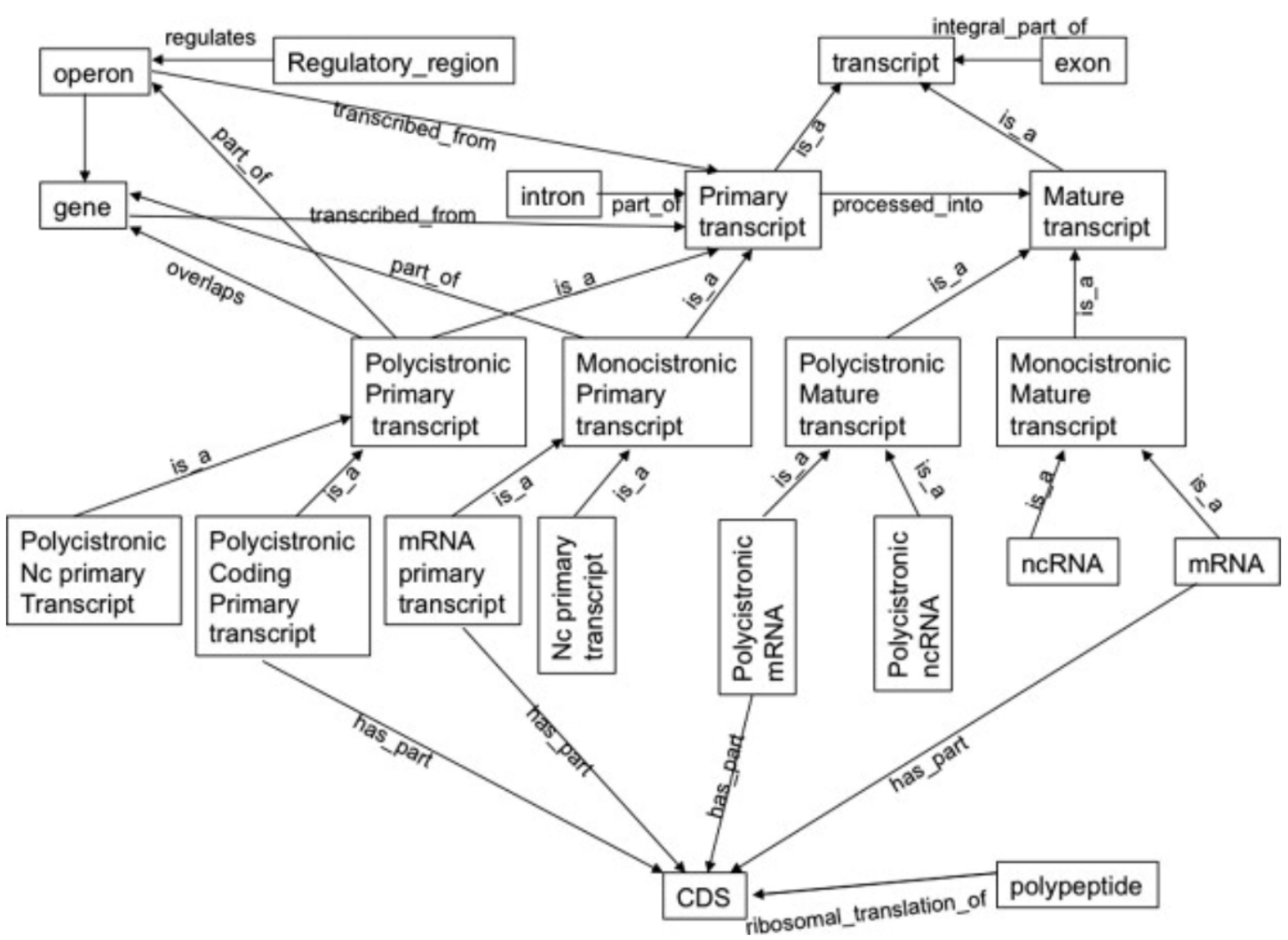
source name	Lung sample from postmortem COVID-19 patient
cell type	Lung Biopsy
strain	NA
subject status	No treatment; >60 years old male COVID-19 deceased patient

How do we fix it?

- Ontologies
- Data standards
  - Prescribed set of fields/terms/formats
- Tools
  - Software to implement standards

How do we fix it?

- Ontologies
  - Hierarchy or trees of controlled vocabularies using standardized terms
  - terms linked using a logical relationship
  - Universal identifiers removes any ambiguity in terms
  - Terms have specific definitions
  - Synonyms



Mungall et al., 2011. Journal of Biomedical Informatics, 44(1): 87-93

Data standards

Standards: ISO 23418:2022

Microbiology of the Food Chain — Whole genome sequencing for typing and genomic characterization of foodborne bacteria — General requirements and guidance

Contextual Data Fields
Sample Collection Lab Contact Information
Geographic Location of Sample Collection
Collection Date
Sample Type
Food Product
Food Processing
Environmental Material
Environmental Location
Collection Device
Collection Method
Microbiology Lab Contact Information
Organism
Strain
Isolate
Serotype
Isolation Media
Isolate Passage History

ISO standard provides tables and annexes to describe…

1. Information about the sample
2. Information about the isolate
3. Information about the sequence

Fields and terms sourced and adapted from:

- Agency documentation
- Public repository submission forms
- Domain expert consultations
- Existing standards and ontologies

ISO slim (package of fields and terms) available:

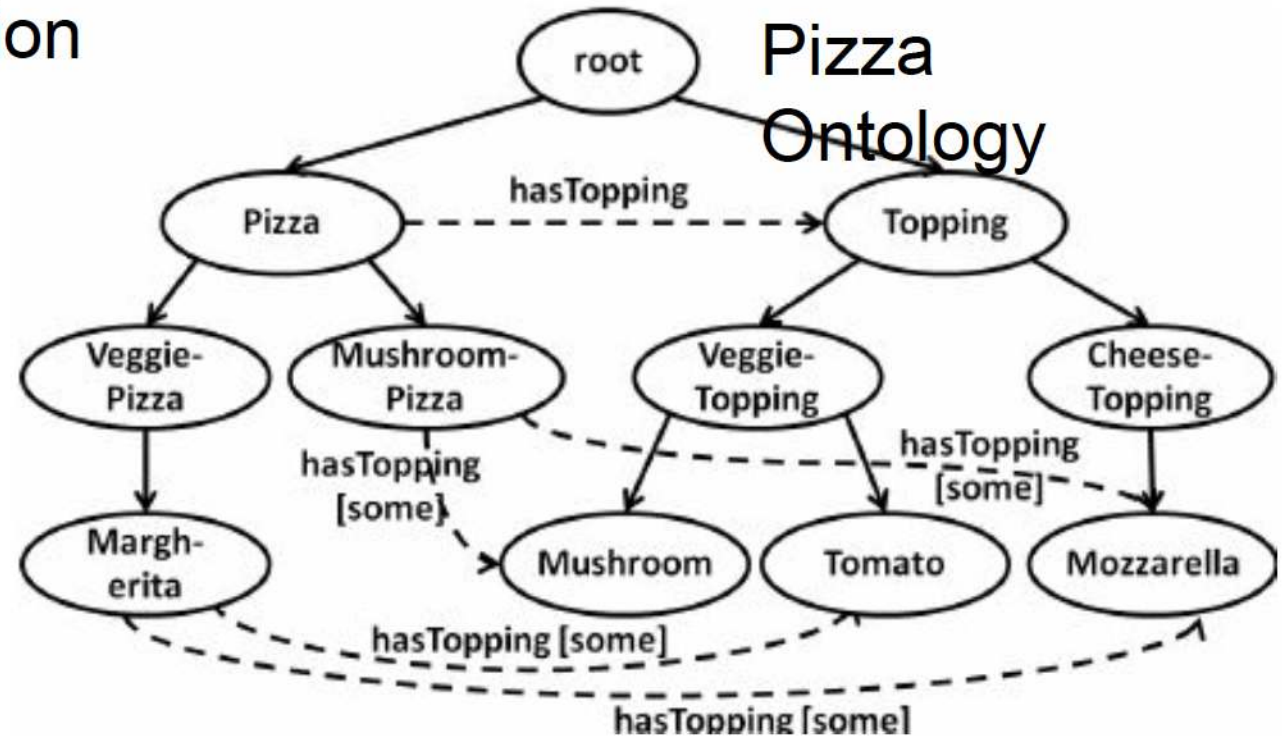
Ontologies: not just lists of terms, but how the terms relate to each

Pizza Data Specification

Pizza types
- Veggie pizza
- Mushroom pizza
- Margherita pizza

Toppings
- Mushroom
- Tomato
- Mozarella

Pizza Ontology



Links particular toppings to particular pizza

Contextual data is critical for interpreting SC2 sequence data.

Sequence data

Contextual data

Sample metadata

Lab results

Clinical/Epi data

Methods

Contextual data (metadata) used for **surveillance** and **outbreak investigations**:

- **characterize** lineages and clusters
- identify variants with **clinical significance**
- correlate genomics trends with **outcomes, risk factors**
- **inform decision making** for public health responses and **monitor effects of interventions**

The SARS-CoV-2 Contextual Data Standard

SARS-CoV-2 Domain Content

- Repository accession numbers and identifiers
- Sample collection and processing
- Host information
- Host exposure information
- Host reinfection information
- Host vaccination information
- Sequencing methods
- Bioinformatics and quality control metrics
- Lineage and variant information
- Pathogen diagnostic testing details
- Provenance and attribution

Data Sources

- Case report forms
- Public repository requirements
- Existing metadata standards
- Literature

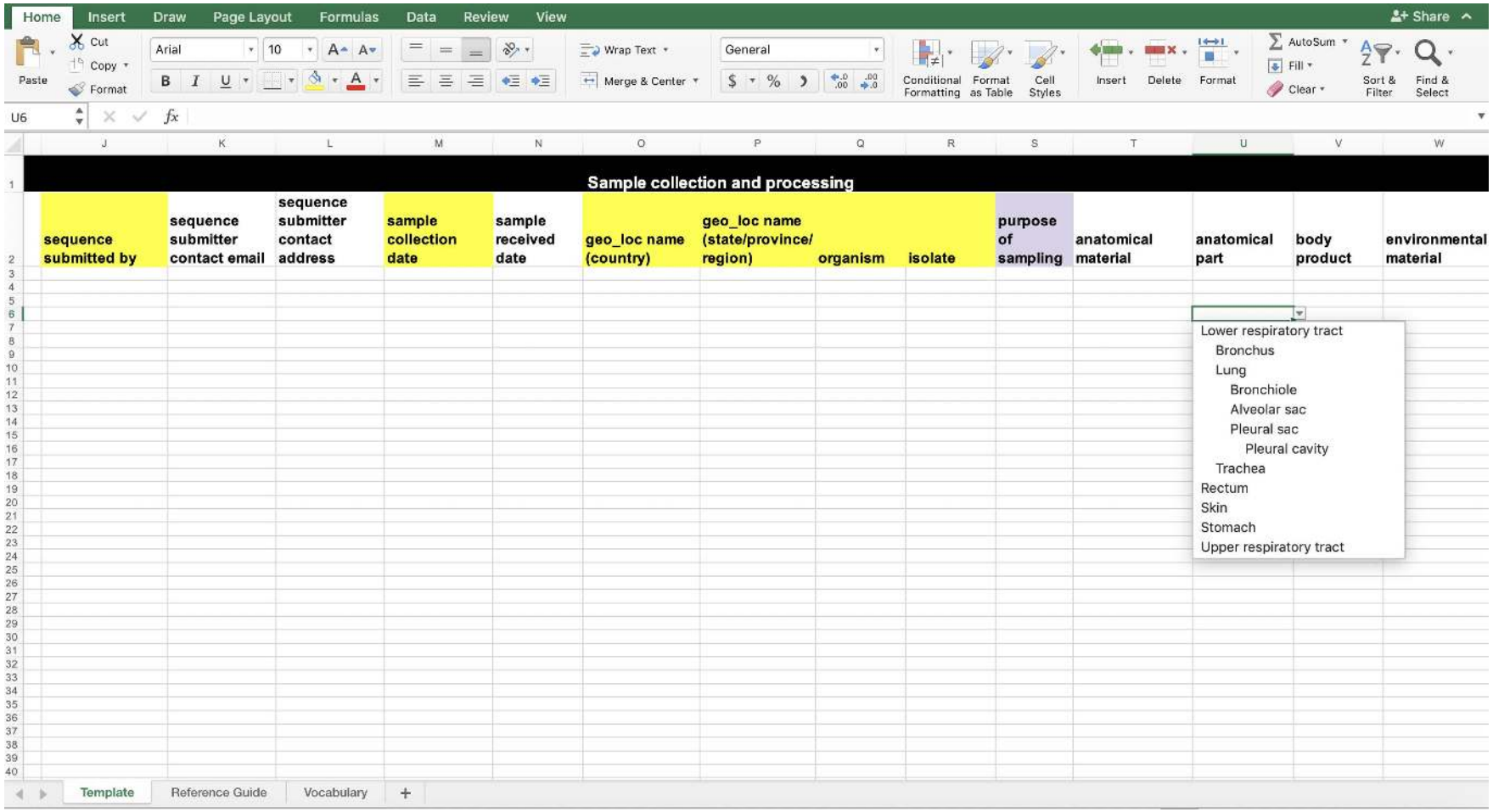
Mapping to Standards

- MIxS 5.0
- MIGS Virus, Host-Associated
- Project/Sample Application Standard
- OBO Foundry Ontologies

PHA4GE SARS-CoV-2 Specification

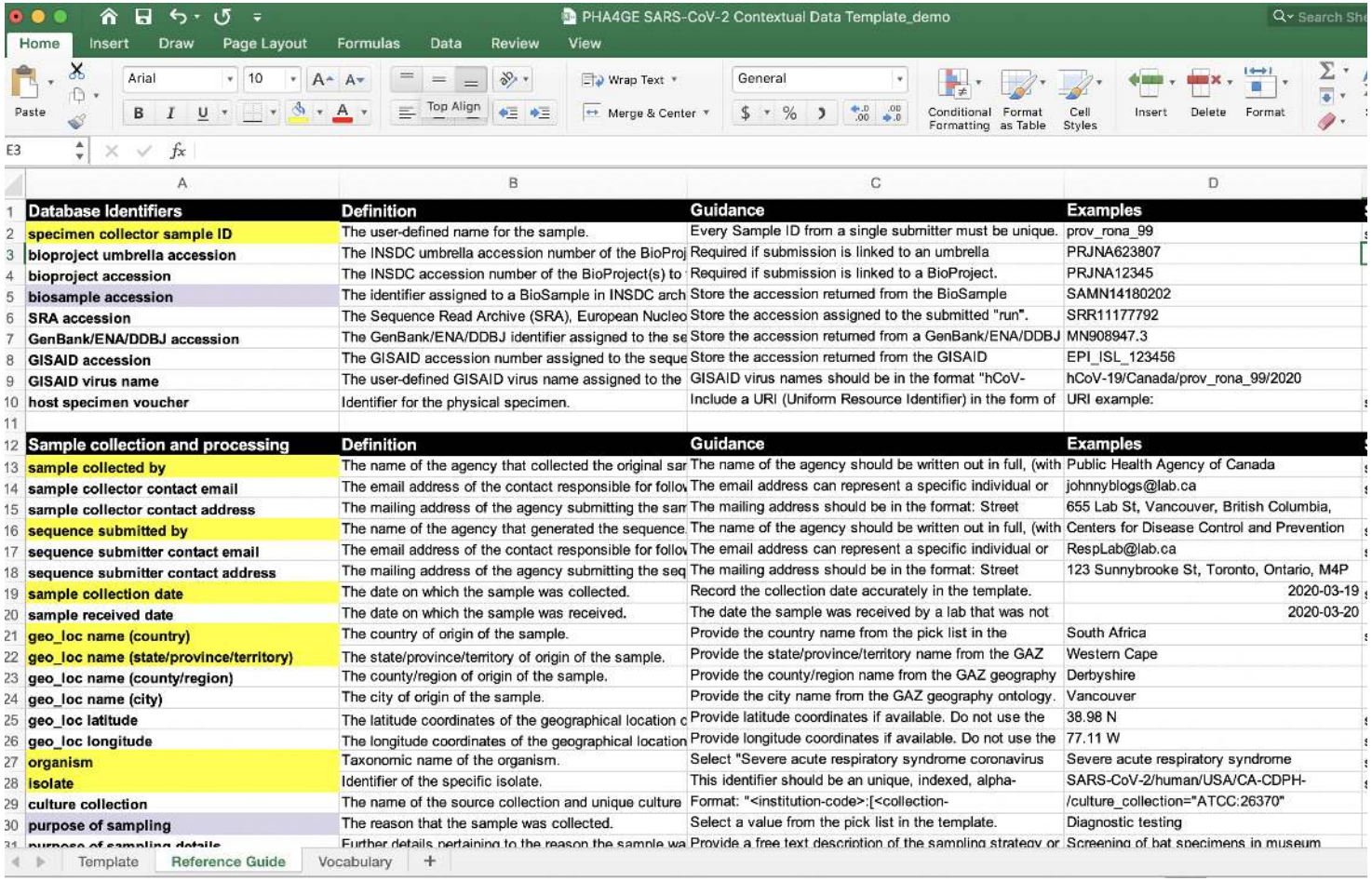
https://github.com/pha4ge/SARS-CoV-2-Contextual-Data-Specification

Putting standards into practice: Template and standard terminology



- **Standardized collection template** (colour-coded, yellow=required, purple=recommended, white=optional)
- **Pick lists**: standardized terms
- **Structured formats** e.g. for dates
- **JSON schema**

Guidance documentation



	A	B	C	D
1	**Database Identifiers**	**Definition**	**Guidance**	**Examples**
2	**specimen collector sample ID**	The user-defined name for the sample.	Every Sample ID from a single submitter must be unique.	prov_rona_99
3	**bioproject umbrella accession**	The INSDC umbrella accession number of the BioProj	Required if submission is linked to an umbrella	PRJNA623807
4	**bioproject accession**	The INSDC accession number of the BioProject(s) to	Required if submission is linked to a BioProject.	PRJNA12345
5	**biosample accession**	The identifier assigned to a BioSample in INSDC arch	Store the accession returned from the BioSample	SAMN14180202
6	**SRA accession**	The Sequence Read Archive (SRA), European Nucleo	Store the accession assigned to the submitted "run".	SRR11177792
7	**GenBank/ENA/DDBJ accession**	The GenBank/ENA/DDBJ identifier assigned to the se	Store the accession returned from a GenBank/ENA/DDBJ	MN908947.3
8	**GISAID accession**	The GISAID accession number assigned to the seque	Store the accession returned from the GISAID	EPI_ISL_123456
9	**GISAID virus name**	The user-defined GISAID virus name assigned to the	GISAID virus names should be in the format "hCoV-	hCoV-19/Canada/prov_rona_99/2020
10	**host specimen voucher**	Identifier for the physical specimen.	Include a URI (Uniform Resource Identifier) in the form of	URI example:
11				
12	**Sample collection and processing**	**Definition**	**Guidance**	**Examples**
13	**sample collected by**	The name of the agency that collected the original sar	The name of the agency should be written out in full, (with	Public Health Agency of Canada
14	**sample collector contact email**	The email address of the contact responsible for follo	The email address can represent a specific individual or	johnnyblogs@lab.ca
15	**sample collector contact address**	The mailing address of the agency submitting the sam	The mailing address should be in the format: Street	655 Lab St, Vancouver, British Columbia,
16	**sequence submitted by**	The name of the agency that generated the sequence.	The name of the agency should be written out in full, (with	Centers for Disease Control and Prevention
17	**sequence submitter contact email**	The email address of the contact responsible for follo	The email address can represent a specific individual or	RespLab@lab.ca
18	**sequence submitter contact address**	The mailing address of the agency submitting the seq	The mailing address should be in the format: Street	123 Sunnybrooke St, Toronto, Ontario, M4P
19	**sample collection date**	The date on which the sample was collected.	Record the collection date accurately in the template.	2020-03-19
20	**sample received date**	The date on which the sample was received.	The date the sample was received by a lab that was not	2020-03-20
21	**geo_loc name (country)**	The country of origin of the sample.	Provide the country name from the pick list in the	South Africa
22	**geo_loc name (state/province/territory)**	The state/province/territory of origin of the sample.	Provide the state/province/territory name from the GAZ	Western Cape
23	**geo_loc name (county/region)**	The county/region of origin of the sample.	Provide the county/region name from the GAZ geography	Derbyshire
24	**geo_loc name (city)**	The city of origin of the sample.	Provide the city name from the GAZ geography ontology.	Vancouver
25	**geo_loc latitude**	The latitude coordinates of the geographical location c	Provide latitude coordinates if available. Do not use the	38.98 N
26	**geo_loc longitude**	The longitude coordinates of the geographical location	Provide longitude coordinates if available. Do not use the	77.11 W
27	**organism**	Taxonomic name of the organism.	Select "Severe acute respiratory syndrome coronavirus	Severe acute respiratory syndrome
28	**isolate**	Identifier of the specific isolate.	This identifier should be an unique, indexed, alpha-	SARS-CoV-2/human/USA/CA-CDPH-
29	**culture collection**	The name of the source collection and unique culture	Format: "<institution-code>:[<collection-	/culture_collection="ATCC:26370"
30	**purpose of sampling**	The reason that the sample was collected.	Select a value from the pick list in the template.	Diagnostic testing
31	**purpose of sampling details**	Further details pertaining to the reason the sample wa	Provide a free text description of the sampling strategy or	Screening of bat specimens in museum

introduced to capture different kinds of anatomical and environmental samples, as well as collection devices and methods. These fields include "anatomical material", "anatomical part", "body product", "environmental material", "environmental site", "collection device", and "collection method". **Populate only the fields that pertain to your sample.** Provide the most granular information allowable according to your organization's data sharing policies.

e.g. nasal swab should be recorded:

host (scientific name)	host (common name)	host disease	anatomical part	collection device
Homo sapiens	Human	COVID-19	Nasopharynx	Swab

e.g. saliva should be recorded:

host (scientific name)	host (common name)	host disease	anatomical material
Homo sapiens	Human	COVID-19	Saliva

e.g. human feces should be recorded:

host (scientific name)	host (common name)	host disease	body product
Homo sapiens	Human	COVID-19	Feces

e.g. sewage from treatment plant should be recorded:

environmental site	environmental material
Sewage Plant	Sewage

e.g. swab of a hospital bed rail should be recorded:

environmental site	environmental material	collection device
Hospital	Bed Rail	Swab

- **Reference guide**: field labels, definitions, guidance, expected values

Mapping Between Formats

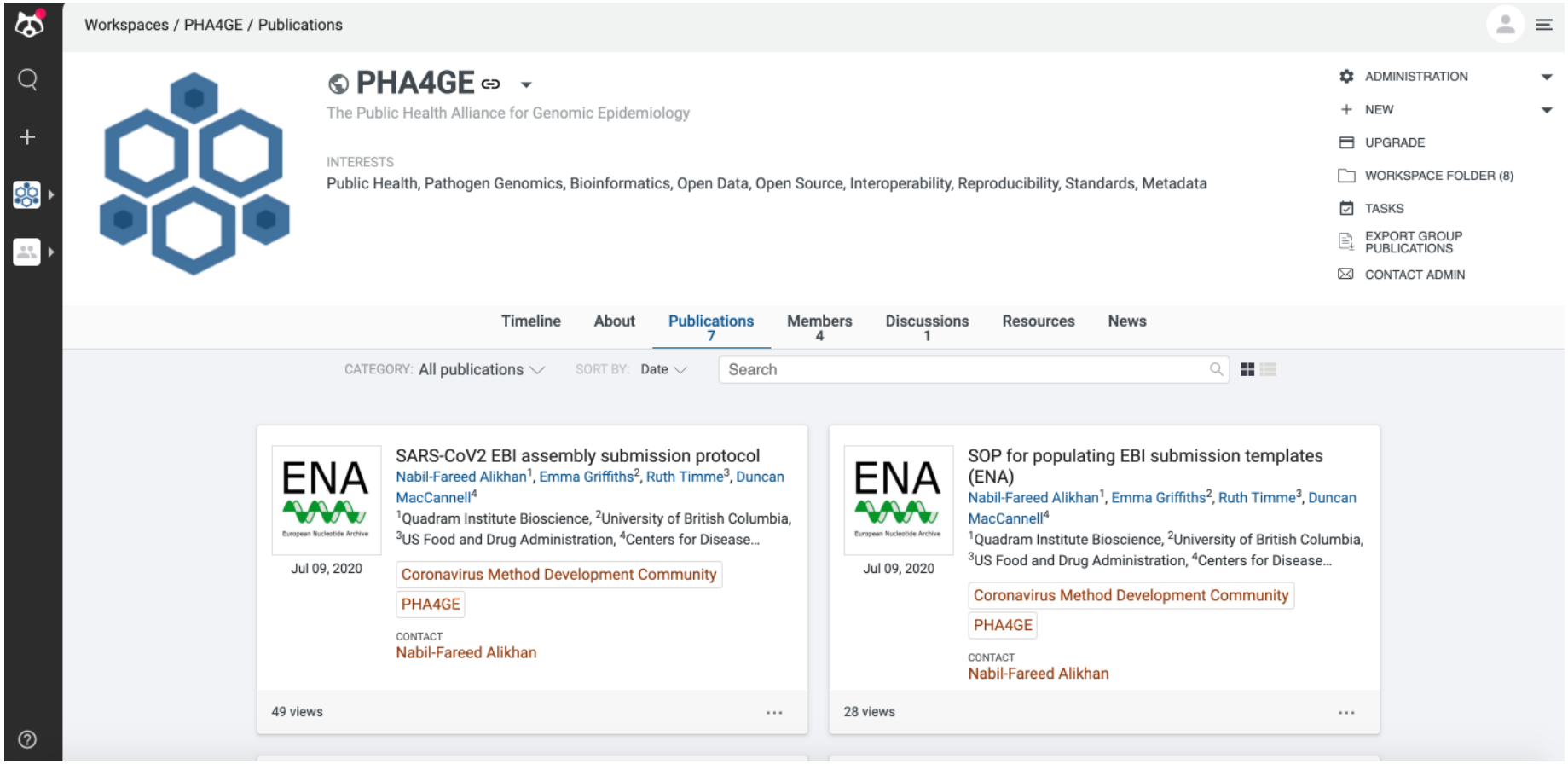
GISAID	ENA Virus Package
Submitter	No equivalent (submit from your account)
FASTA filename	file_name (See Experiment metadata)
Virus name	isolate
Type	tax_id (See Experiment metadata)
Passage details/history	No equivalent
Collection date	collection date
Location	geographic location (country and/or sea)
Host	host common name/host scientific name
Gender	host sex
Patient age	host age
Patient status	host health state
Sequencing technology	instrument_model (See Experiment metadata)
Originating lab	collecting institution
Address	collecting institution
Submitting lab	No equivalent (submit from your account)
Address	No equivalent (submit from your account)
No equivalent	host subject id

GISAID Virus name:
hCoV-19/Country/ISO regional code-Identifier
hCoV-19/Country/**un**-Identifier/year
e.g. hCoV-19/CANADA/BC-ABCD1234/2021

NCBI Isolate:
SARS-CoV-2/host/country(short)/sampleID/da
e.g. SARS-CoV-2/human/CAN/ABCD1234/20

*remember, even if something is "required", you can always provide a null value if you need to
e.g. Missing, Not Applicable, Not Collected

Protocols to mobilize harmonized data



- **7 public repository submission protocols** (GISAID, NCBI, EMBL-EBI) on **Protocols.io**
- **PHA4GE-adapted submission forms**
- **instructional videos**

Different repositories have different fields, but PHA4GE helps standardize what goes into those fields

https://www.protocols.io/workspaces/pha4ge

Data stewardship: oversight and practices to ensure data is **accessible, usable, safe, trusted.**

Privacy protection (sharing):

- **Public trust essential**, loss of trust has consequences (protection, transparency)
- De-identified data (**no names/addresses**)
- Be careful of 1) geographical granularity, 2) small case numbers in defined geo_loc/time, 3) combinations of fields
- **Track identifiers** (chain of custody), but personal health IDs/sample IDs may be considered PHII
- **Consult privacy officer** (jurisdictional policies, national legislation)

Security & Quality:

- Provenance, methods (rich details) □ **attribution, auditability, reproducibility (track methods), accountability**
- Contextual data may require storage with **higher security** than seq data
- Errors **corrected**, **update** as required

Types of contextual data critical for surveillance/ genomic epidemiology (what you can most likely share)

- **Geo_loc (at least country, preferably state/province) – sample collection**
- **Sample collection date (to the day)**
- **Attribution: Who collected sample, who sequenced it**
- **Methods: instrument (platform & model), consensus sequence software, coverage**

- Sampling strategy (random sampling, targeted sampling, outbreak, research)
- Demographics: age/sex (gender)
- Sample type
- Host
- Quality indicators (e.g. Ct values)

- Vaccination
- Exposures
- Travel history
- Hospitalization
- Outcomes

Slide content acknowledgement

- Slide 4-11: data carpentries genomics workshop

- Slide 12-26: Emma Griffiths bioinformaticsdotca.github.io