SARS-CoV-2 genomics contextual data curation and stewardship

Emma Griffiths, PhD

Chair, PHA4GE Data Structures Working Group

Hsiao Public Health Bioinformatics Lab

Faculty of Health Sciences, Simon Fraser University

Vancouver, Canada

Public Health Alliance for Genomic Epidemiology

Outline

1. Challenges associated with genomics contextual data

2. Data standards: solutions for harmonization

3. PHA4GE SARS-CoV-2 specification
   - Resources for putting standards into practice

4. Data stewardship

5. Wrap up & links

Public Health Alliance for Genomic Epidemiology

Contextual data is critical for interpreting the 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**

Public Health Alliance for
Genomic Epidemiology

Harmonizing fields of data is challenging.

A field by any other name does NOT smell as sweet…

SPECIMENSOURCE_1
Isolation
host_tissue_sampled
Source

The labs mean
“**sample type**”

Differences in labels,
Same meaning

Computer doesn’t recognize these as the same thing

Source

The lab means
“**submitting lab**”

Same label,
Different meaning

Computer doesn’t recognize these as different

…so, you can’t just combine fields of data.

Public Health Alliance for
Genomic Epidemiology

Harmonizing information inside fields is challenging.

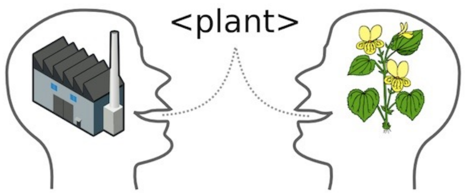
Free text = ☹

Sever Acute
Repiratry Sickness

Errors

SC2

Short hand



Semantic ambiguity

Date:
2021-04-26
April 26, 2021
26-Apr-2021

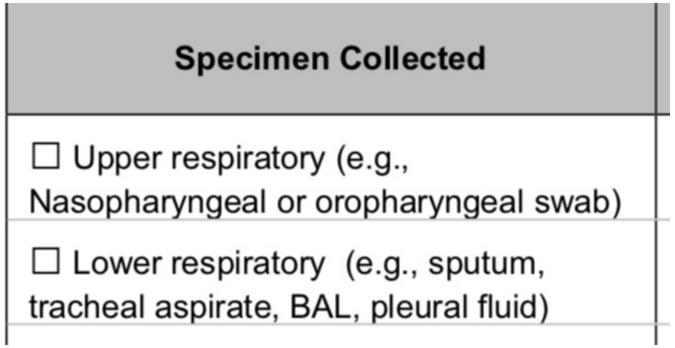
Formats

Inconsistently collected

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

- Need to share data: **within** organization, with **trusted partners**, with **international** agencies/**public** repositories
- Data structure variability in local databases propagates to public repositories

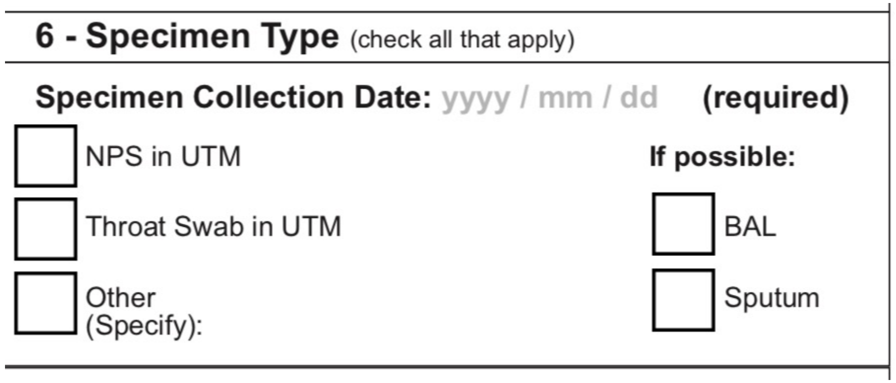
Private databases:



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

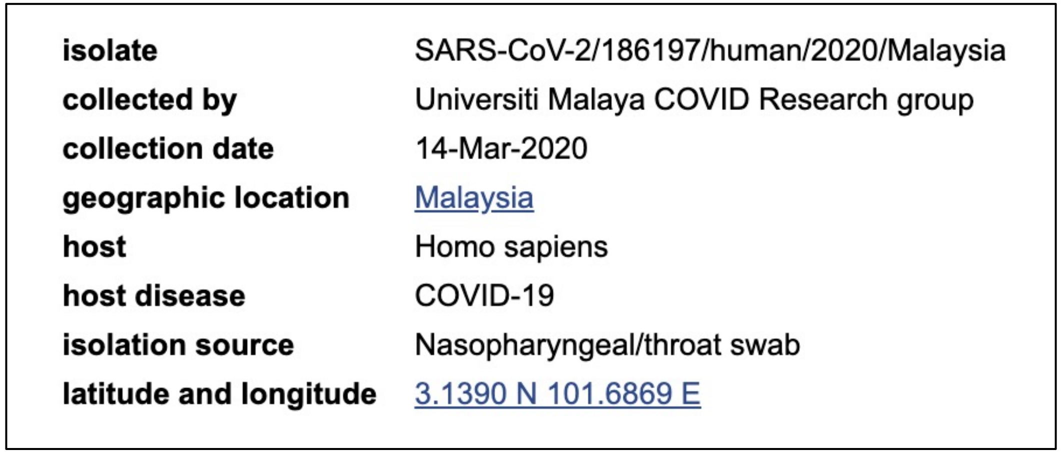
☐ Other (Specify):

If possible:

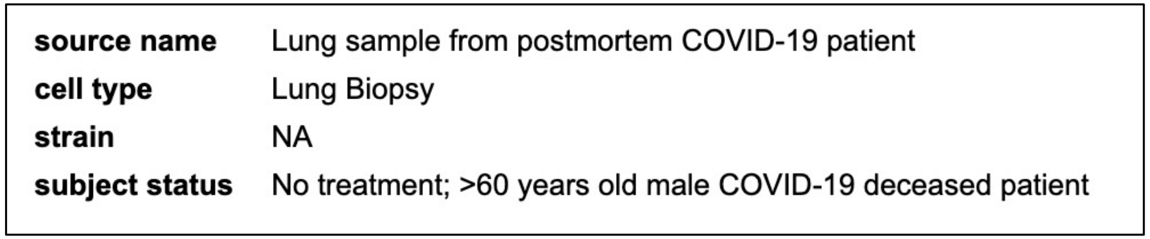
☐ BAL

☐ 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

Different data structures make information less interoperable and more difficult to integrate.

That means you need to spend more time and resources to clean/re-structure information before you can use it.

Best practices for data **management/stewardship/structure** are critical parts of SARS-CoV-2 sequencing and analyses.

Public Health Alliance for
Genomic Epidemiology

Data standards: Solutions for standardizing contextual data

Data dictionaries:

- **Fields, flat list of terms, formats**
- Data dictionaries fulfill particular purposes
- Organization or project specific project (ease of use)

e.g. Pizza Type

- Cheese pizza
- Meat Lover's Pizza
- Veggie Pizza

Minimum Information Checklists:

- **Prescribed fields to describe something (in a particular context)**
- created by authoritative source
- core fields common between checklists (e.g. collected by, sample collection date, sample ID)
- specific packages provide fields for specific contexts (e.g. air, human gut)
- commonly used by international sequence repositories

e.g. MIAP (Min Info About a Pizza)

Pizza Type:

Topping:

Crust:

Sauce:

Public Health Alliance for Genomic Epidemiology

Ontology, A Way of Structuring Information

Ontologies aim to represent truth. *Is this universal?*

- Controlled (standardized) vocabulary
- Hierarchy (granularity)
- Logic
- Definitions and unique IDs (disambiguation)

e.g. Veggie Pizza (FoodOn:1234)

A type of pizza that is topped with only vegetable and cheese toppings.

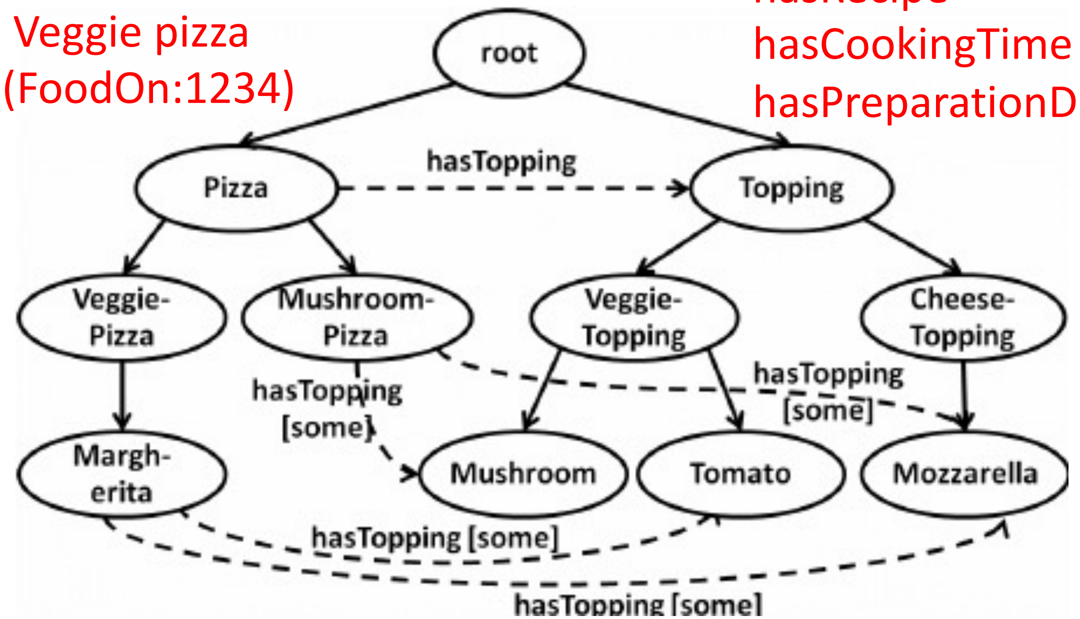
e.g. Margarita – synonym of Margherita?

IDs distinguish between pizza and cocktail

- Synonyms (facilitates mapping)
- Prepares data for more variety of analyses

Veggie pizza (FoodOn:1234)

hasRecipe
hasCookingTime
hasPreparationDate

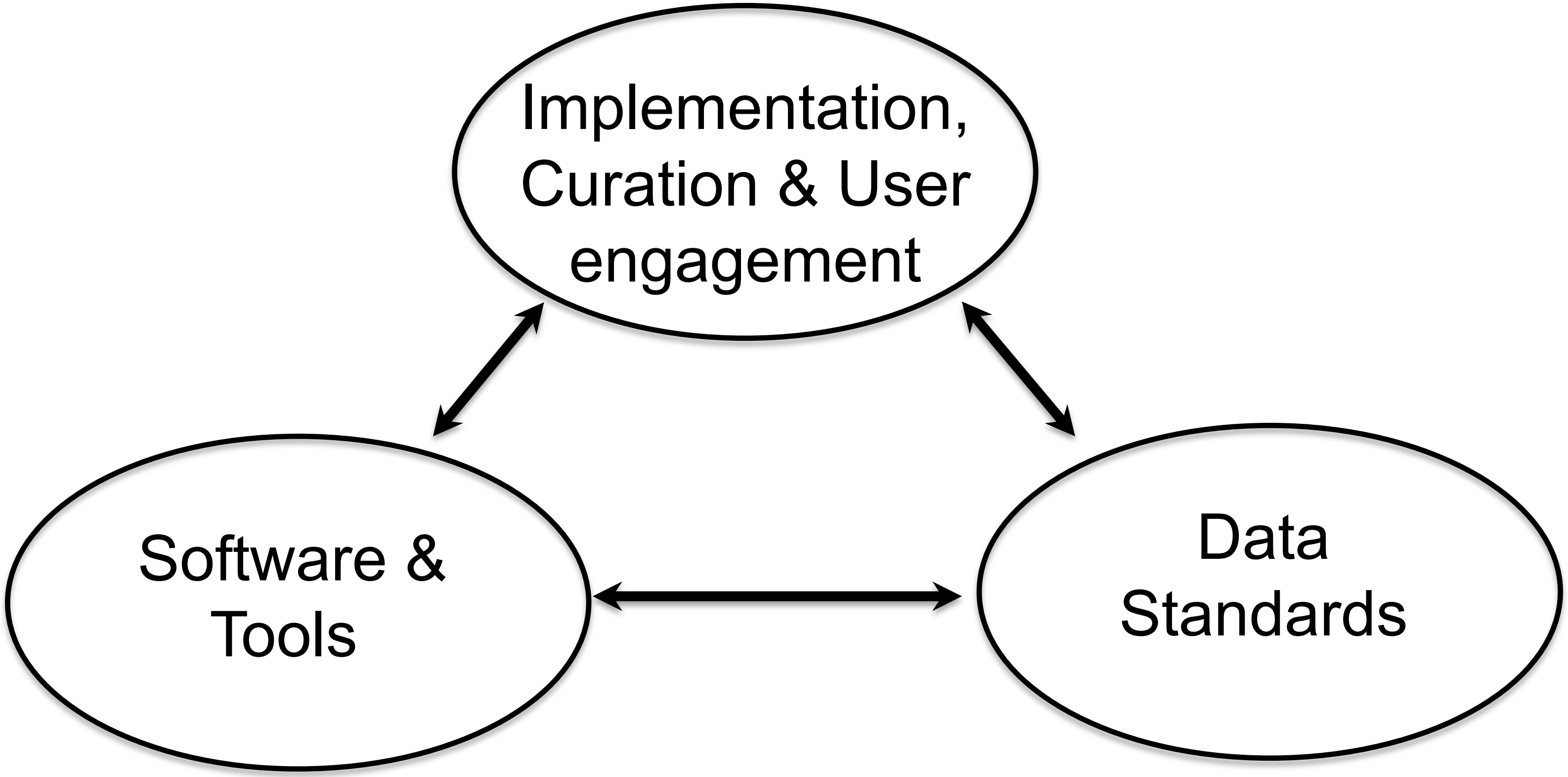


Margherita pizza (FoodOn:1240)
synonym: Margarita pizza

Margarita cocktail (FoodOn:2376)
Synonym: Margarita

Public Health Alliance for Genomic Epidemiology

Contextual data ecosystem: putting standards into practice



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

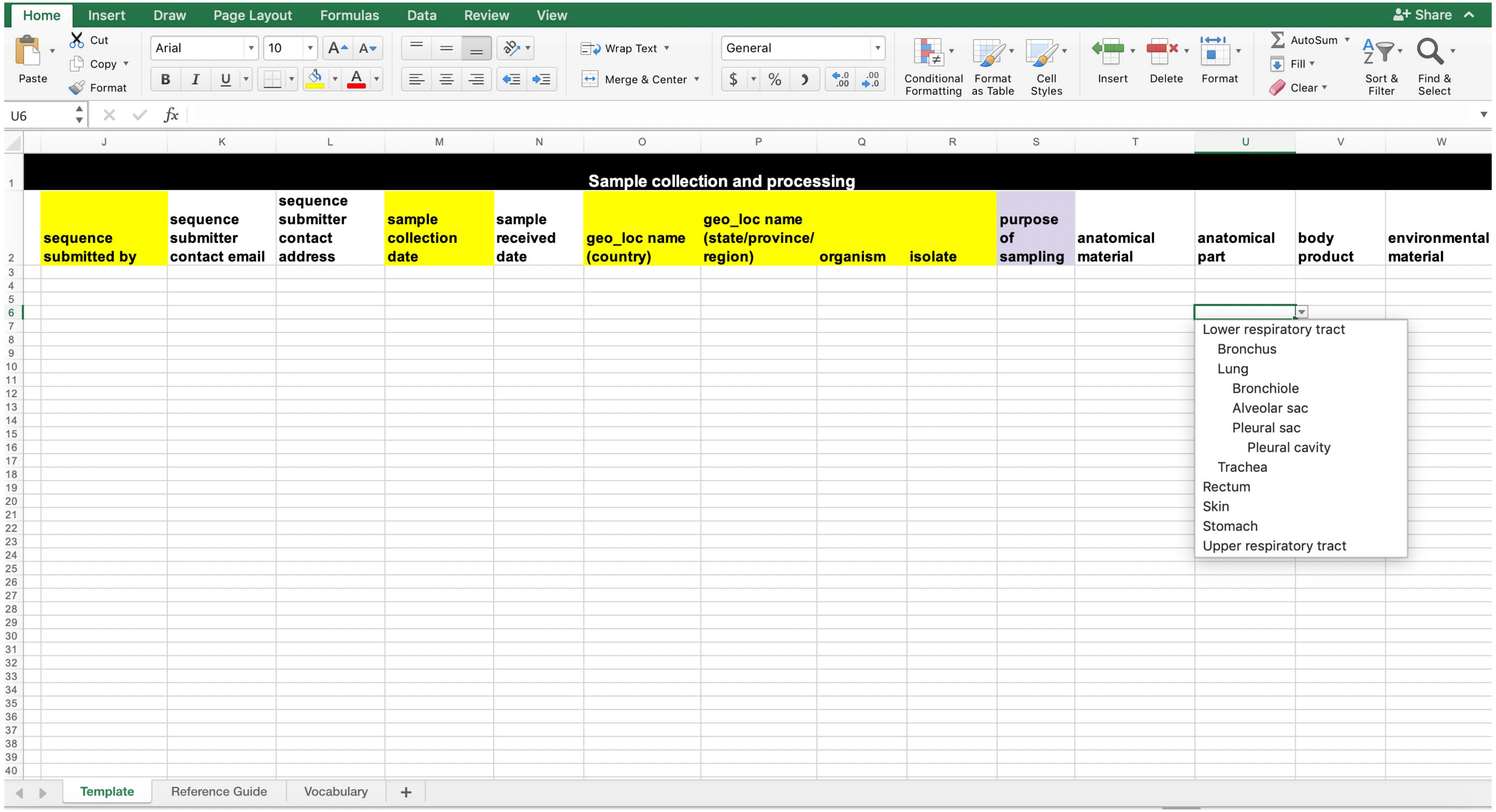
Mapping to Standards

- 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

Public Health Alliance for Genomic Epidemiology

Putting standards into practice: Template and standard terminology



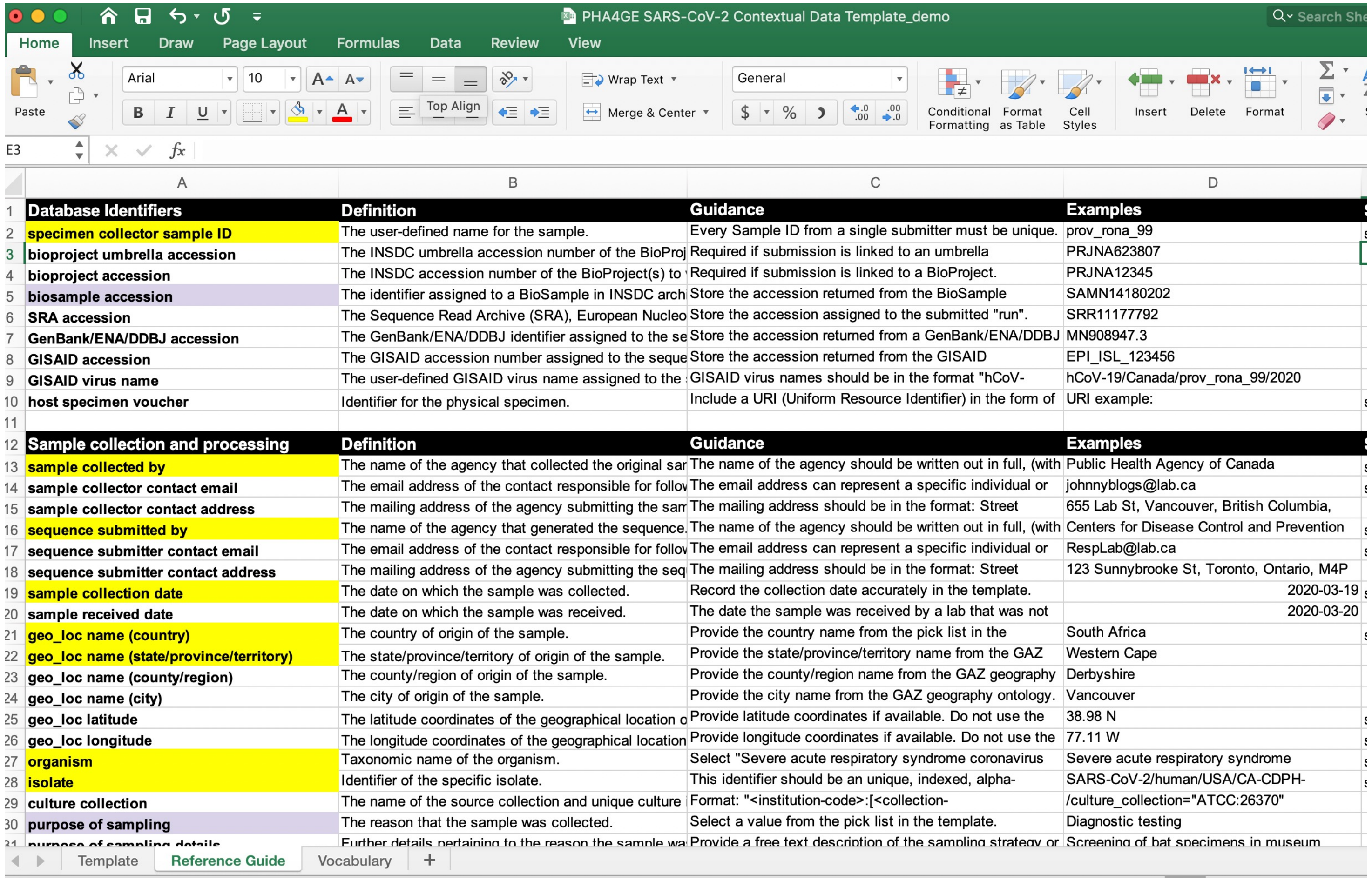
	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	Sample collection and processing													
2	sequence submitted by	sequence submitter contact email	sequence submitter contact address	sample collection date	sample received date	geo_loc name (country)	geo_loc name (state/province/ region)	organism	isolate	purpose of sampling	anatomical material	anatomical part	body product	environmental material

- Lower respiratory tract
  - Bronchus
  - Lung
    - Bronchiole
    - Alveolar sac
    - Pleural sac
      - Pleural cavity
  - Trachea
- Rectum
- Skin
- Stomach
- Upper respiratory tract

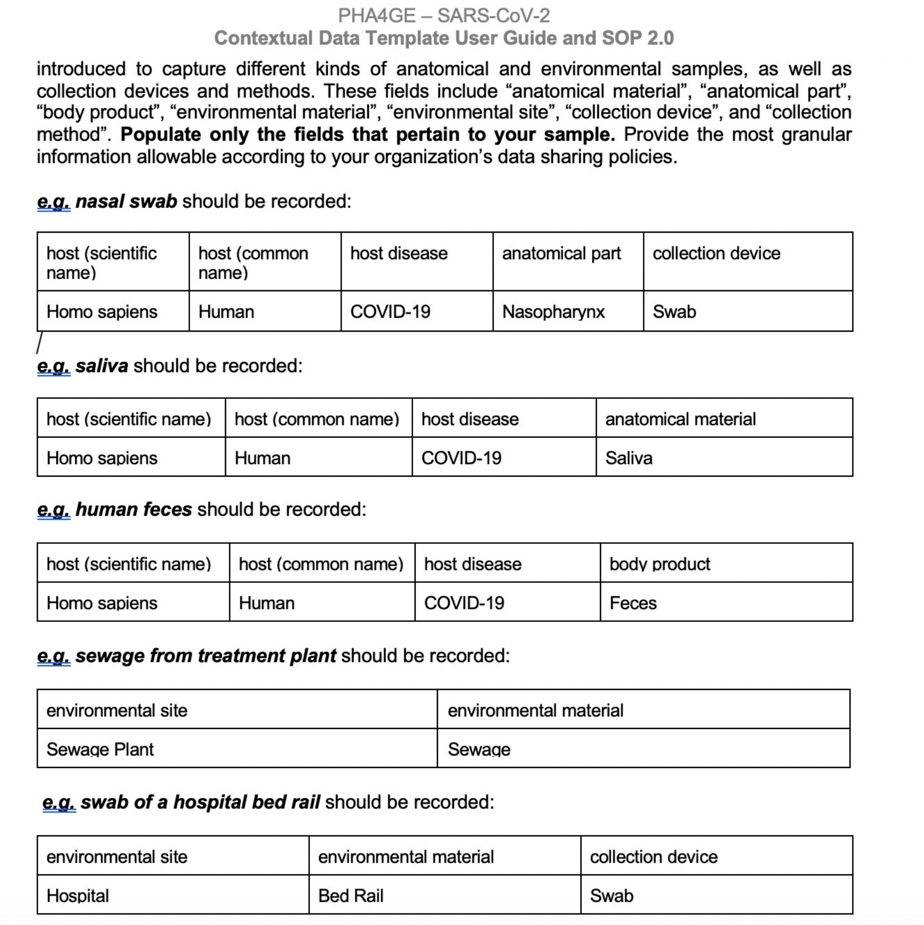
- **Standardized collection template** (colour-coded, yellow=required, purple=recommended, white=optional)
- **Pick lists**: standardized terms
- **Structured formats** e.g. for dates
- **JSON schema**

Public Health Alliance for Genomic Epidemiology

Guidance documentation



Database Identifiers	Definition	Guidance	Examples
specimen collector sample ID	The user-defined name for the sample.	Every Sample ID from a single submitter must be unique.	prov_rona_99
bioproject umbrella accession	The INSDC umbrella accession number of the BioProj	Required if submission is linked to an umbrella	PRJNA623807
bioproject accession	The INSDC accession number of the BioProject(s) to	Required if submission is linked to a BioProject.	PRJNA12345
biosample accession	The identifier assigned to a BioSample in INSDC arch	Store the accession returned from the BioSample	SAMN14180202
SRA accession	The Sequence Read Archive (SRA), European Nucleo	Store the accession assigned to the submitted "run".	SRR11177792
GenBank/ENA/DDBJ accession	The GenBank/ENA/DDBJ identifier assigned to the se	Store the accession returned from a GenBank/ENA/DDBJ	MN908947.3
GISAID accession	The GISAID accession number assigned to the seque	Store the accession returned from the GISAID	EPI_ISL_123456
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
host specimen voucher	Identifier for the physical specimen.	Include a URI (Uniform Resource Identifier) in the form of	URI example:
Sample collection and processing	**Definition**	**Guidance**	**Examples**
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
sample collector contact email	The email address of the contact responsible for follov	The email address can represent a specific individual or	johnnyblogs@lab.ca
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,
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
sequence submitter contact email	The email address of the contact responsible for follov	The email address can represent a specific individual or	RespLab@lab.ca
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
sample collection date	The date on which the sample was collected.	Record the collection date accurately in the template.	2020-03-19
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
geo_loc name (country)	The country of origin of the sample.	Provide the country name from the pick list in the	South Africa
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
geo_loc name (county/region)	The county/region of origin of the sample.	Provide the county/region name from the GAZ geography	Derbyshire
geo_loc name (city)	The city of origin of the sample.	Provide the city name from the GAZ geography ontology.	Vancouver
geo_loc latitude	The latitude coordinates of the geographical location o	Provide latitude coordinates if available. Do not use the	38.98 N
geo_loc longitude	The longitude coordinates of the geographical location	Provide longitude coordinates if available. Do not use the	77.11 W
organism	Taxonomic name of the organism.	Select "Severe acute respiratory syndrome coronavirus	Severe acute respiratory syndrome
isolate	Identifier of the specific isolate.	This identifier should be an unique, indexed, alpha-	SARS-CoV-2/human/USA/CA-CDPH-
culture collection	The name of the source collection and unique culture	Format: "<institution-code>:[<collection-	/culture_collection="ATCC:26370"
purpose of sampling	The reason that the sample was collected.	Select a value from the pick list in the template.	Diagnostic testing
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



PHA4GE – SARS-CoV-2
Contextual Data Template User Guide and SOP 2.0

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
- **SOP**: how to curate contextual data

Public Health Alliance for Genomic Epidemiology

PHA4GE standard quick FAQ

Do I have to fill in the whole thing?

NO! *Only use the parts you need. We've highlighted the most important bits.*

Is this just for human/clinical samples?

NO! *It's for ALL samples.*

Do I have to share all my contextual data?

NO! *It's all up to you!*

What happens if your pick lists don't have the term I want?

1. Get in touch with us!

2. SOP shows you how to find a standardized term.

Worked Examples

- **State X reference lab** in **Country Y** has a **border testing program**.
- A **36 year old female** from **Canada visiting Country Y** tests positive for **SARS-CoV-2**.
- An **NP swab** was collected on **March 16 2021**.
- The sample was sequenced by **State X reference lab** using an amplicon strategy according to the **ARTIC V4** protocol.
- The consensus genome was generated **iVar 1.3.2**.

specimen collector sample ID: ABCD1234
sample collected by: State X Reference Laboratory
sequence submitted by: State X Reference Laboratory
sample collection date: 2021-03-16
geo_loc name (country): Country Y
geo_loc name (state/province/territory): State X
purpose of sampling: Diagnostic testing
purpose of sampling details: Not Provided
purpose of sequencing: Travel-associated surveillance
purpose of sequencing details: border testing program
organism: Severe Acute Respiratory Syndrome Coronavirus 2

***Null values:**

Missing
Not Applicable
Not Collected
Not Provided
Restricted Access

anatomical part: Nasopharynx (NP)
collection device: Swab
host (scientific name): Homo sapiens
host age: 36
host age unit: year
host gender: Female
host residence geo_loc name (country): Canada
travel history: individual travelled directly from Canada
amplicon pcr primer scheme: ARTIC V4
amplicon size: 400bp
consensus sequence software name: iVar
consensus sequence software version: 1.3.2

Public Health Alliance for
Genomic Epidemiology

Worked Examples

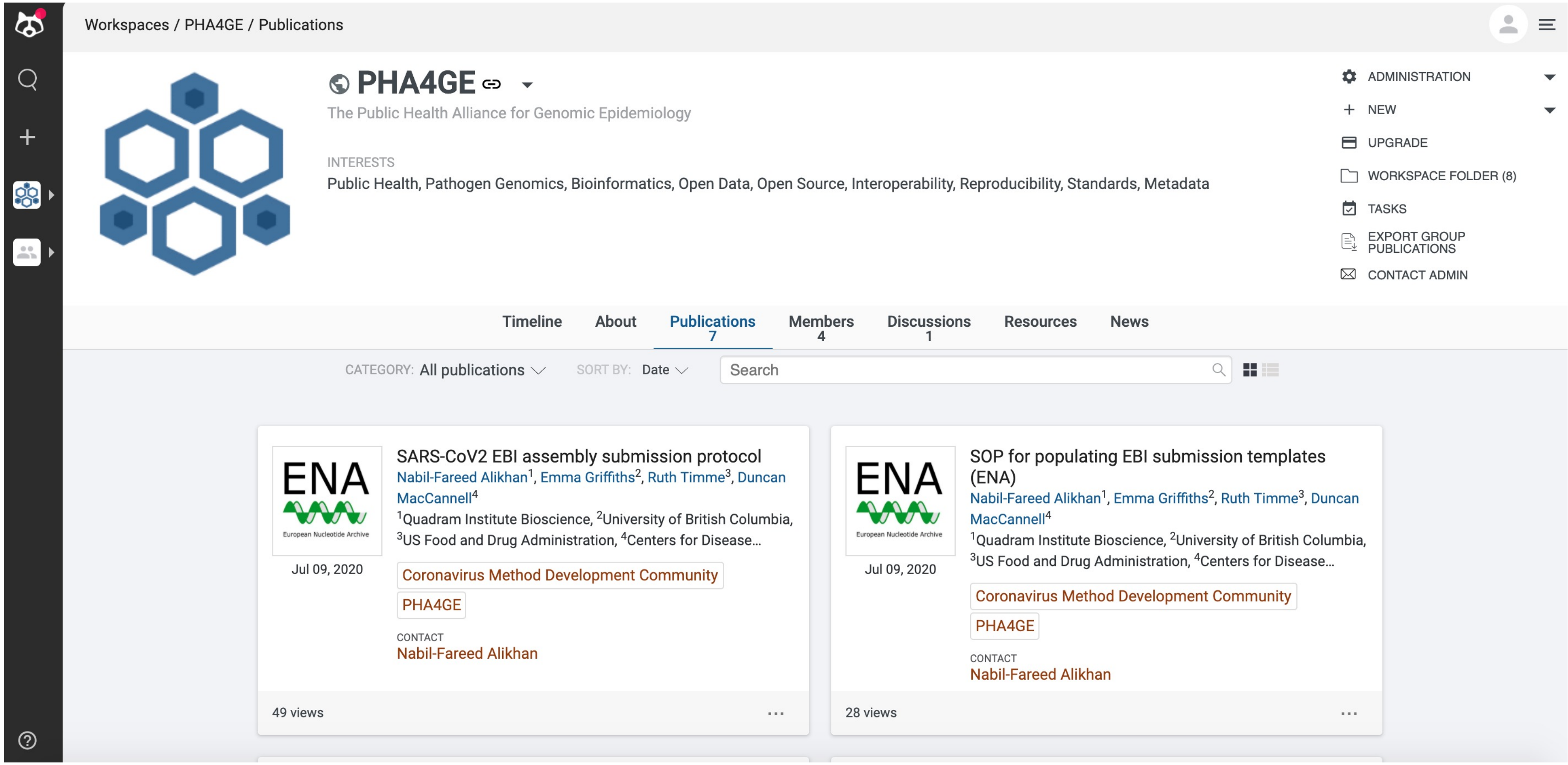
- **State X reference lab** in **Country Y** in investigating an **outbreak** at **Hospital Z**.
- A previously **vaccinated** individual (**one dose AZ, Feb 1 2021**) tests positive for SARS-CoV-2, with a diagnostic PCR **CT value of 23 (E gene)**.
- The sample was sequenced by State X reference lab using an **Illumina MiSeq** (**Illumina prep kit**).
- The sequence data was **filtered, processed and dehosted** using the **SIGNAL pipeline**. The consensus genome was generated using **FreeBayes 1.2.3**.
- Using **Pangolin**, the genome was identified as a **Delta VOC**.

specimen collector sample ID: ABCD1234
sample collected by: State X Reference Laboratory
sequence submitted by: State X Reference Laboratory
geo_loc name (country): Country Y
geo_loc name (state/province/territory): State X
purpose of sampling: Diagnostic testing
purpose of sequencing: Cluster/Outbreak investigation
purpose of sequencing details: outbreak at Hospital Z
organism: Severe Acute Respiratory Syndrome Coronavirus 2
host (scientific name): Homo sapiens
exposure setting: Hospital
host role: Patient
host vaccination status: Partially vaccinated
vaccine name: AstraZeneca COVISHIELD COVID-19 vaccine (ChAdOx1-S)
number of vaccine doses received: 1
first dose vaccination date: 2021-02-01

sequencing instrument: Illumina MiSeq
library preparation kit: Illumina Prep Kit
raw sequence data processing method: https://github.com/phac-nml/covid-19-signal-nml
dehosting method: https://github.com/phac-nml/covid-19-signal-nml
consensus sequence software name: FreeBayes
consensus sequence software version: 1.2.3
lineage/clade name: B.1.617.2
lineage/clade analysis software name: Pangolin
lineage/clade analysis software version: 3.1.4
variant designation: Variant of Concern (VOC)
gene name 1: E gene (orf4)
diagnostic pcr Ct value 1: 23

Public Health Alliance for Genomic Epidemiology

Protocols to mobilize harmonized data



PHA4GE

The Public Health Alliance for Genomic Epidemiology

INTERESTS

Public Health, Pathogen Genomics, Bioinformatics, Open Data, Open Source, Interoperability, Reproducibility, Standards, Metadata

ENA European Nucleotide Archive

Jul 09, 2020

SARS-CoV2 EBI assembly submission protocol

Nabil-Fareed Alikhan[1], Emma Griffiths[2], Ruth Timme[3], Duncan MacCannell[4]

[1]Quadram Institute Bioscience, [2]University of British Columbia, [3]US Food and Drug Administration, [4]Centers for Disease...

Coronavirus Method Development Community

PHA4GE

CONTACT

Nabil-Fareed Alikhan

49 views

ENA European Nucleotide Archive

Jul 09, 2020

SOP for populating EBI submission templates (ENA)

Nabil-Fareed Alikhan[1], Emma Griffiths[2], Ruth Timme[3], Duncan MacCannell[4]

[1]Quadram Institute Bioscience, [2]University of British Columbia, [3]US Food and Drug Administration, [4]Centers for Disease...

Coronavirus Method Development Community

PHA4GE

CONTACT

Nabil-Fareed Alikhan

28 views

- **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

Public Health Alliance for Genomic Epidemiology

Data transformation tools: The DataHarmonizer

- Tool for data entry and validation
- Spreadsheet-style text editor application (PHA4GE template)
- Picklists, data structure, validation, export
- Guidance (reference guide), SOP

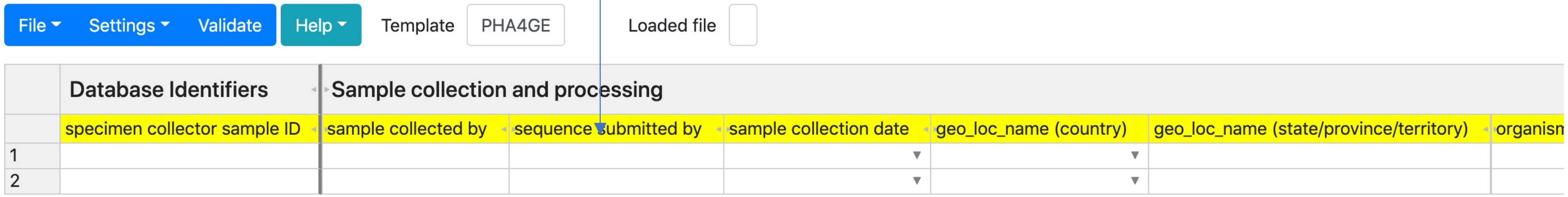
View all fields
View required fields
Move to desired field
Automate column fill

Validate
(check for errors
or missing info)

Learn your
way around
the system

Double click on
field labels for
guidance on how
to fill them

Save
Open existing file
Export to chosen format



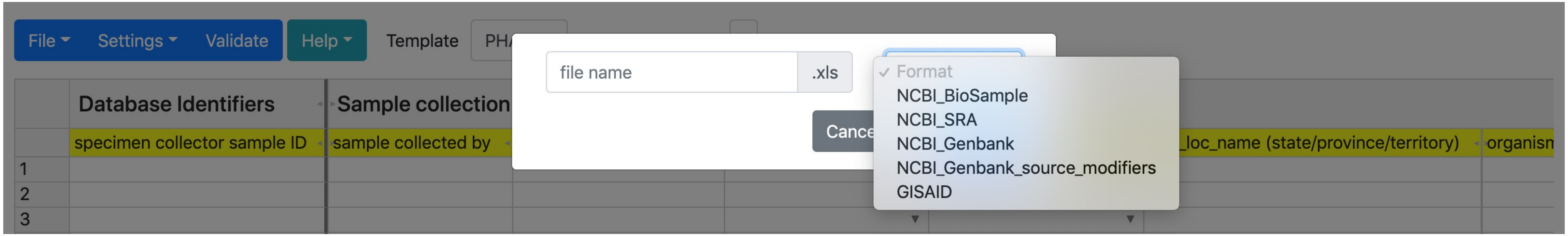
Template PHA4GE

Loaded file

	Database Identifiers	Sample collection and processing					
	specimen collector sample ID	sample collected by	sequence submitted by	sample collection date	geo_loc_name (country)	geo_loc_name (state/province/territory)	organisn
1				▾	▾		
2				▾	▾		

https://github.com/Public-Health-Bioinformatics/DataHarmonizer/wiki/PHA4GE-SARS-CoV-2-Template

Data transformation tools: The DataHarmonizer



- Enter data once, export in different submission formats i.e. GISAID and NCBI (BioSample, SRA, GenBank)

- Complementary to **multiSub** (interchange between submission formats)
- https://github.com/maximilianh/multiSub

PUBLIC HEALTH BIOINFORMATICS

https://github.com/Public-Health-Bioinformatics/DataHarmonizer/wiki/PHA4GE-SARS-CoV-2-Template

Public Health Alliance for Genomic Epidemiology

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
- Age/gender often shareable, sharing of clinical/epi information may be restricted at individual level (e.g. hospitalization, exposures)
- Track identifiers (chain of custody), but personal health IDs may be considered PHII
- Consult privacy officer (jurisdictional policies)
- ***See curation SOP for details…***

Contextual data can provide critical information in investigations of life or death situations. Please consider when evaluating privacy concerns (see GA4GH's Human rights argument for prioritizing sharing).

https://www.ga4gh.org/genomic-data-toolkit/regulatory-ethics-toolkit/framework-for-responsible-sharing-of-genomic-and-health-related-data/

Public Health Alliance for Genomic Epidemiology

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

FAIR stewardship best practices (Wilkinson et al (2016), Nature):

- Findable, Accessible, Interoperable, Reusable
- Standards help preserve integrity and meaning of data (now and in future) for YOU and others
- Machine-actionable → indexed, searchable
- Formal, broadly applicable language (data standards)

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

Public Health Alliance for Genomic Epidemiology

Summary: What can the PHA4GE spec do for you?

✓ 1. Data is more interpretable by humans and computers

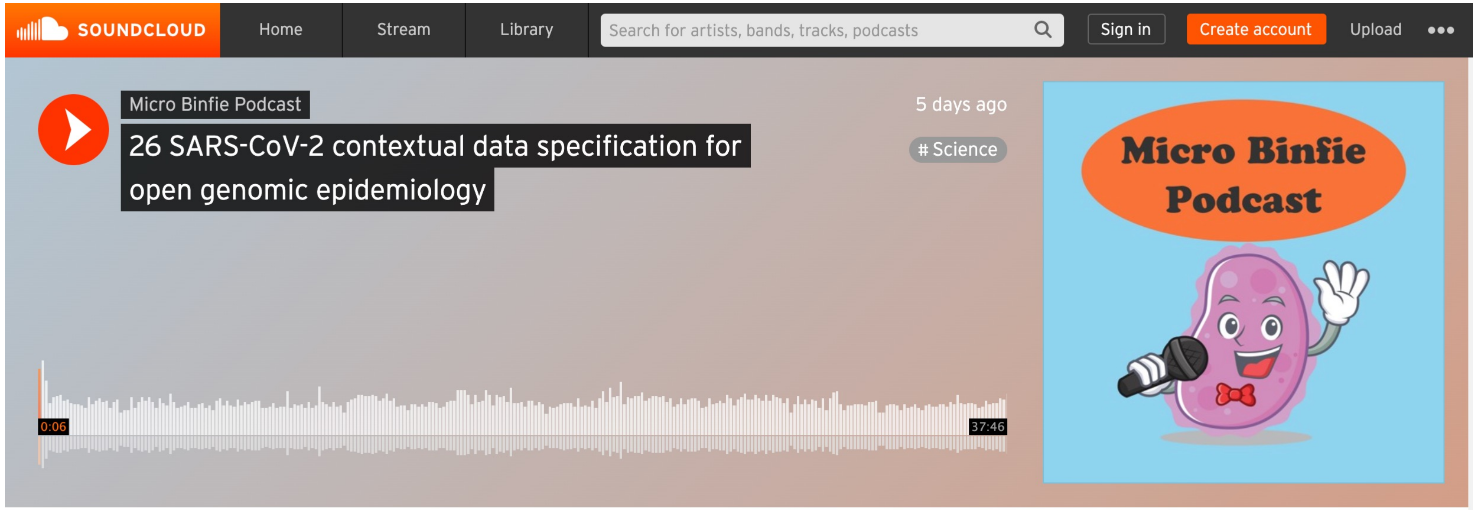
✓ 2. Future-proof contextual data

✓ 3. Harmonize and integrate data across labs/databases (interoperability)

✓ 4. Tools

Public Health Alliance for Genomic Epidemiology

Learn more…

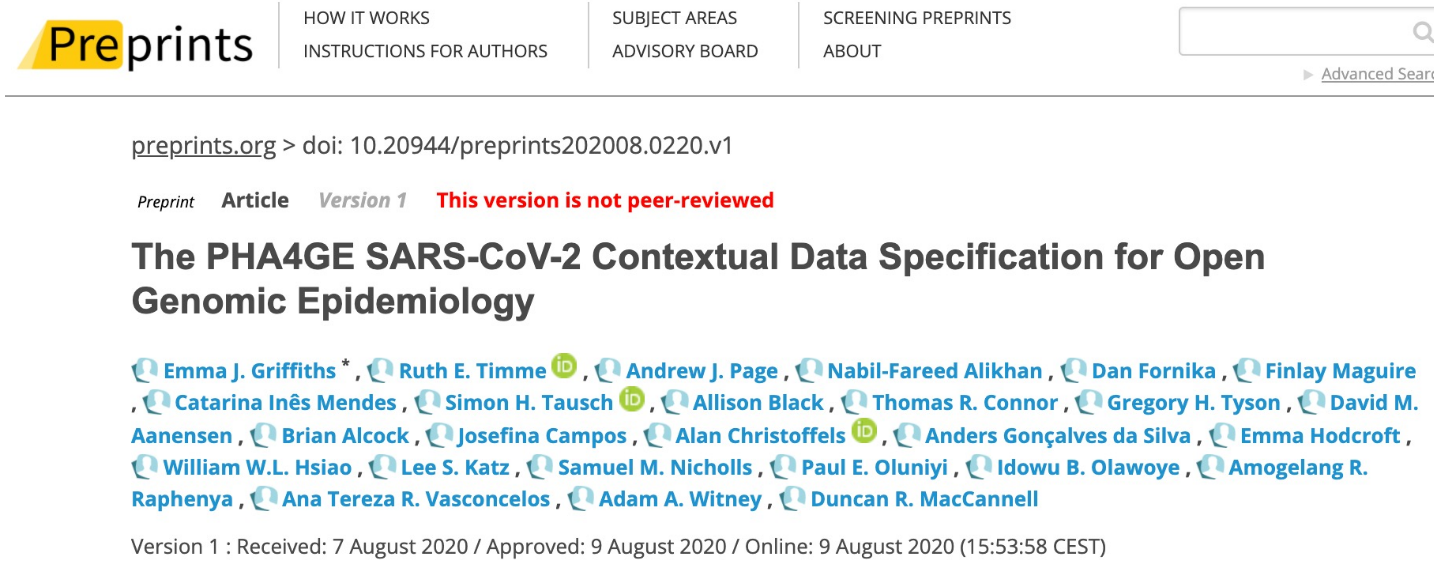


Micro Binfie Podcast

26 SARS-CoV-2 contextual data specification for open genomic epidemiology

https://soundcloud.com/microbinfie/26-sars-cov-2-metadata#t=0:00

Listen to episode 26 Micro Binfie podcast



preprints.org > doi: 10.20944/preprints202008.0220.v1

Preprint **Article** *Version 1* **This version is not peer-reviewed**

The PHA4GE SARS-CoV-2 Contextual Data Specification for Open Genomic Epidemiology

Emma J. Griffiths *, Ruth E. Timme, Andrew J. Page, Nabil-Fareed Alikhan, Dan Fornika, Finlay Maguire, Catarina Inês Mendes, Simon H. Tausch, Allison Black, Thomas R. Connor, Gregory H. Tyson, David M. Aanensen, Brian Alcock, Josefina Campos, Alan Christoffels, Anders Gonçalves da Silva, Emma Hodcroft, William W.L. Hsiao, Lee S. Katz, Samuel M. Nicholls, Paul E. Oluniyi, Idowu B. Olawoye, Amogelang R. Raphenya, Ana Tereza R. Vasconcelos, Adam A. Witney, Duncan R. MacCannell

Version 1 : Received: 7 August 2020 / Approved: 9 August 2020 / Online: 9 August 2020 (15:53:58 CEST)

https://www.preprints.org/manuscript/202008.0220/v1

Read our preprint

Update coming out soon!

Public Health Alliance for Genomic Epidemiology

Special thanks to…

Data Structures Team

Ana Ribeiro de Vasconcelos
Josefina Campos
Idowu Olawoye
Paul Oluniyi
Adam Witney
Andrew Page
David Aanensen
Ines Mendes
Emma Hodcroft
Simon Tausch
Allison Black
Ruth Timme
Greg Tyson
Mike Feldgarden
Lee Katz
Brian Alcock
Amos Raphenya
Finlay Maguire
Dan Fornika
Duncan MacCannell

Specification Contributors & Partners

Nabil-Fareed Alikhan
Alan Christoffels
Will Hsiao
Sam Nicholls
Tom Connor
Anders Gonçalves da Silva
Dominique Anderson
Danny Park
CDC TOAST Team

Steering Committee & Secretariat

Jamie Southgate
Alecia Naidu
Rangarirai Matima
Nawaal Nacerodien
Peter van Heusden
Kevin Libuit
Nicola Mulder
Nicki Tiffin

Find the spec package:
https://github.com/pha4ge/SARS-CoV-2-Contextual-Data-Specification

Find us:
https://www.pha4ge.org
@pha4ge
datastructures@pha4ge.org

Public Health Alliance for Genomic Epidemiology