Fitting data together to fight the COVID-19 pandemic

The PHA4GE SARS-CoV-2 genomic surveillance contextual data specification package

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 associate with contextual data
2. PHA4GE SARS-COV-2 standard – what's in it and its benefits.
3. Quick FAQ.
4. Wrap up and 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

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	**If possible:**	
☐	Throat Swab in UTM	☐	BAL
☐	Other (Specify):	☐	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

Why the PHA4GE SARS-CoV-2 standard?

- Different contextual data standards out there (public repo submission requirements, MIxS, NIAID Sample & Application Standard)
- Harmonized by experts involved in national sequencing efforts & standards development public health focus
- Being adopted around the world

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

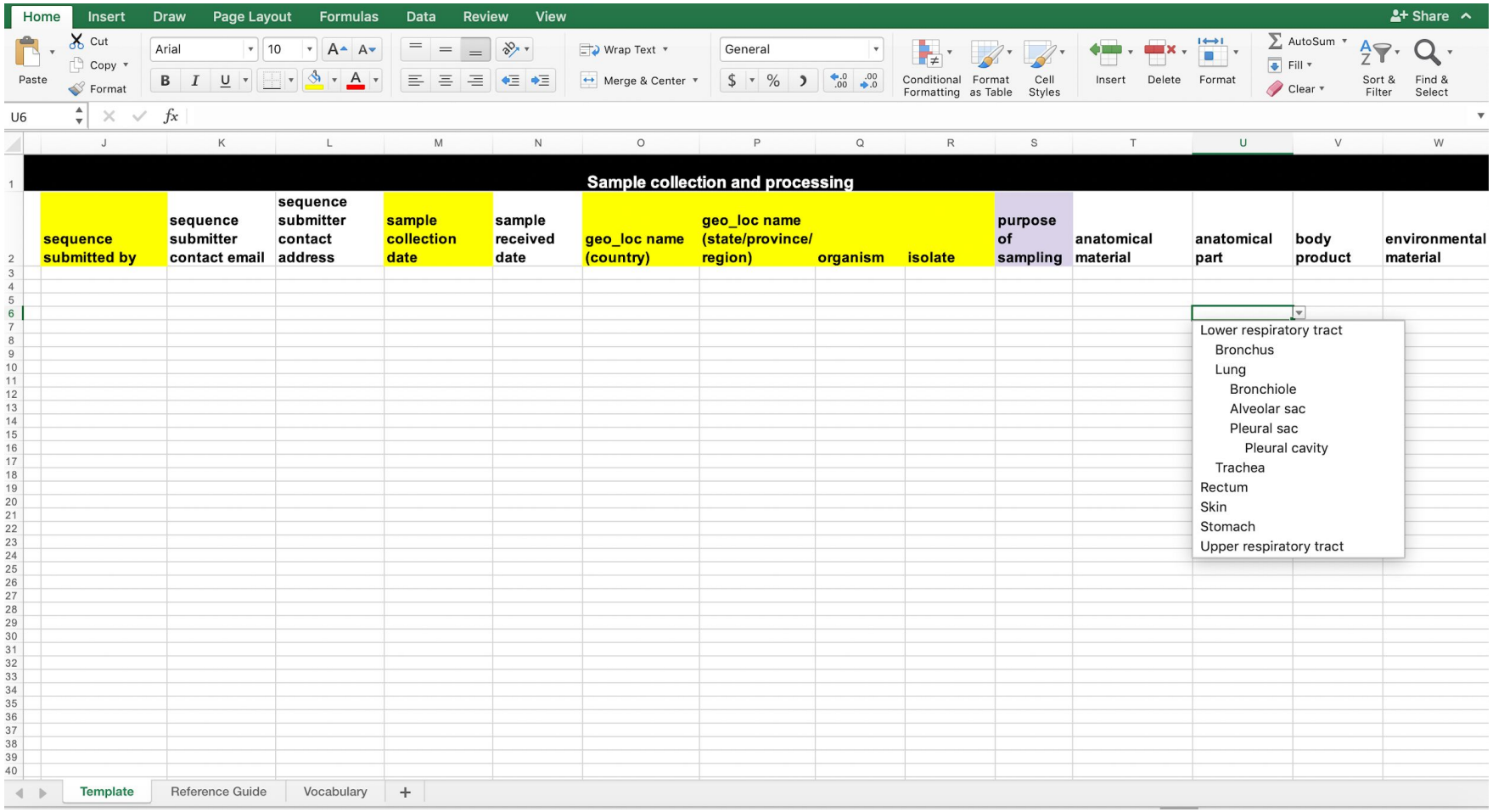
Mapping to Standards

- 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

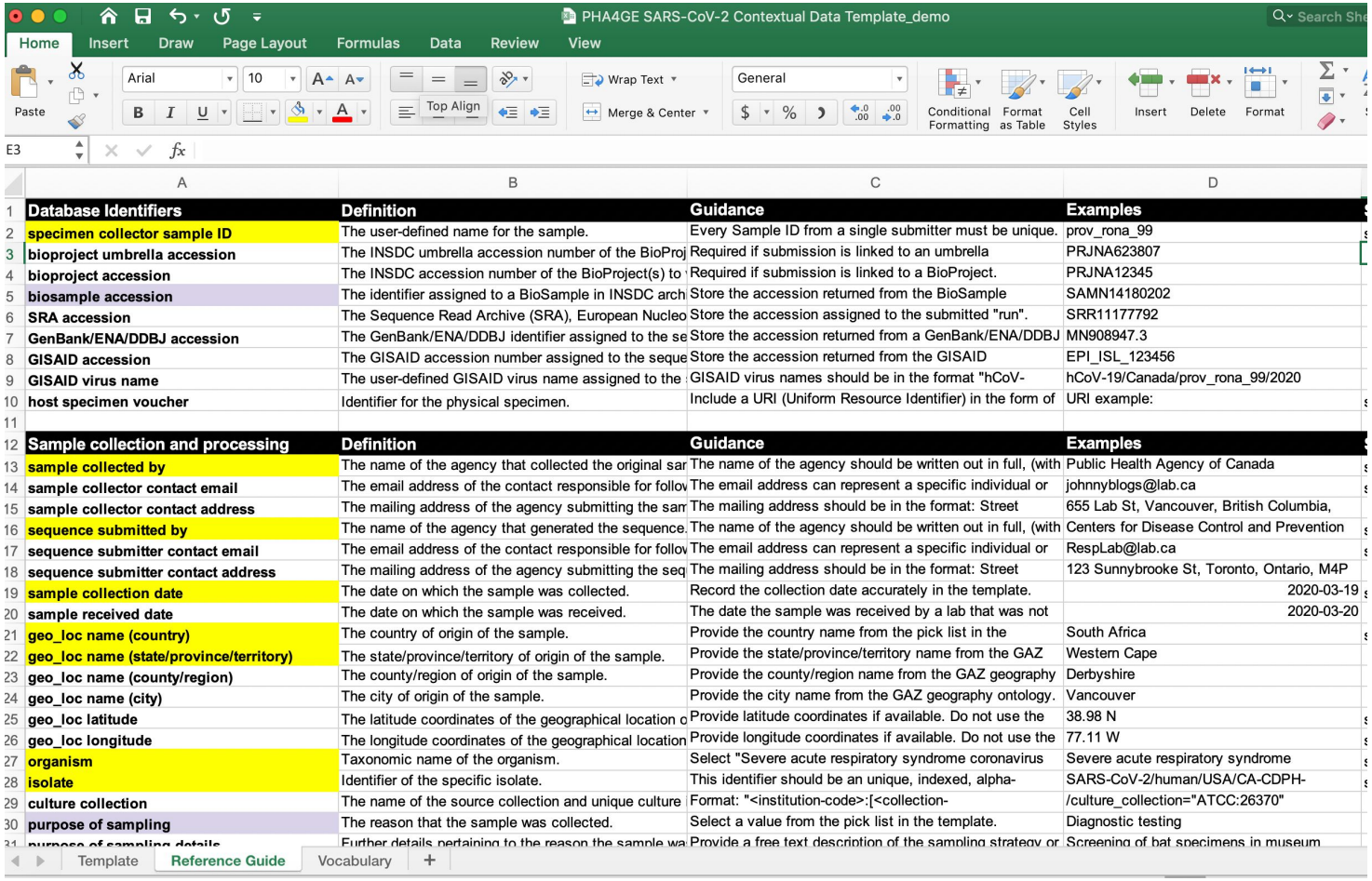
Template and standard terminology



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

Public Health Alliance for Genomic Epidemiology

Guidance documentation

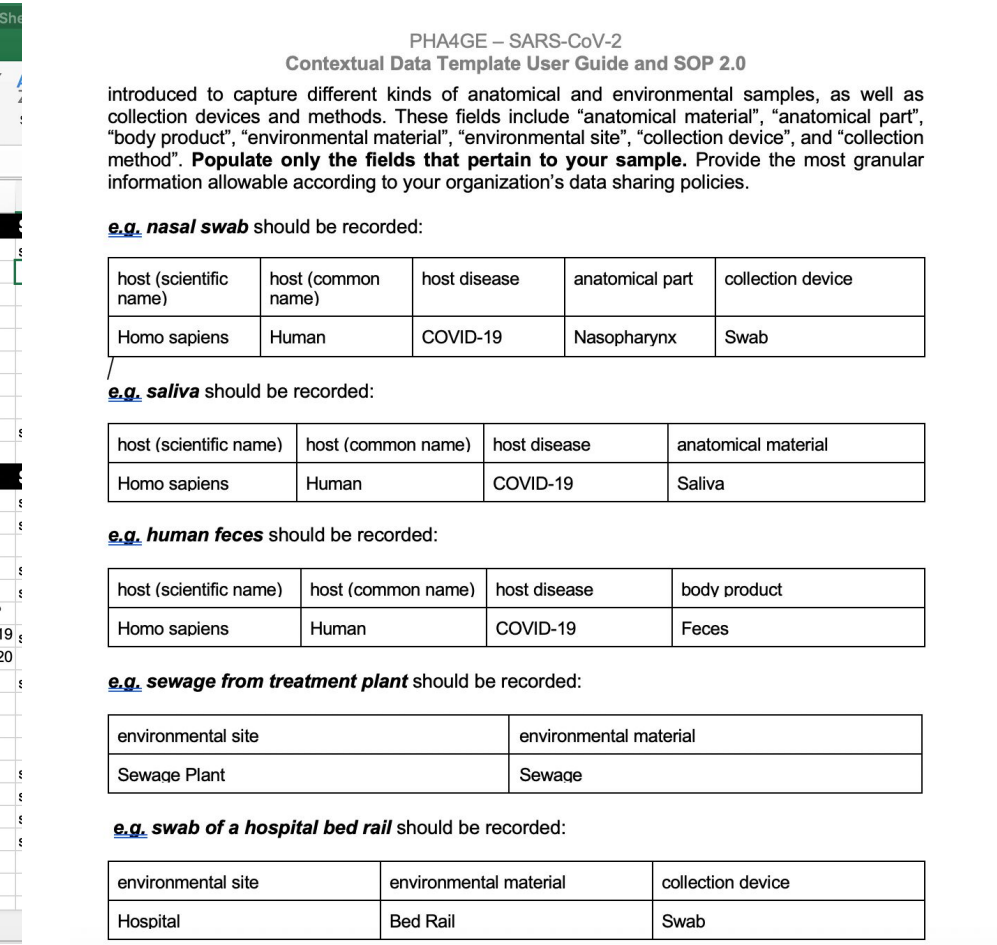


PHA4GE SARS-CoV-2 Contextual Data Template_demo

Home | Insert | Draw | Page Layout | Formulas | Data | Review | View

	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 follov	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 follov	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 o	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

Template | Reference Guide | Vocabulary



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

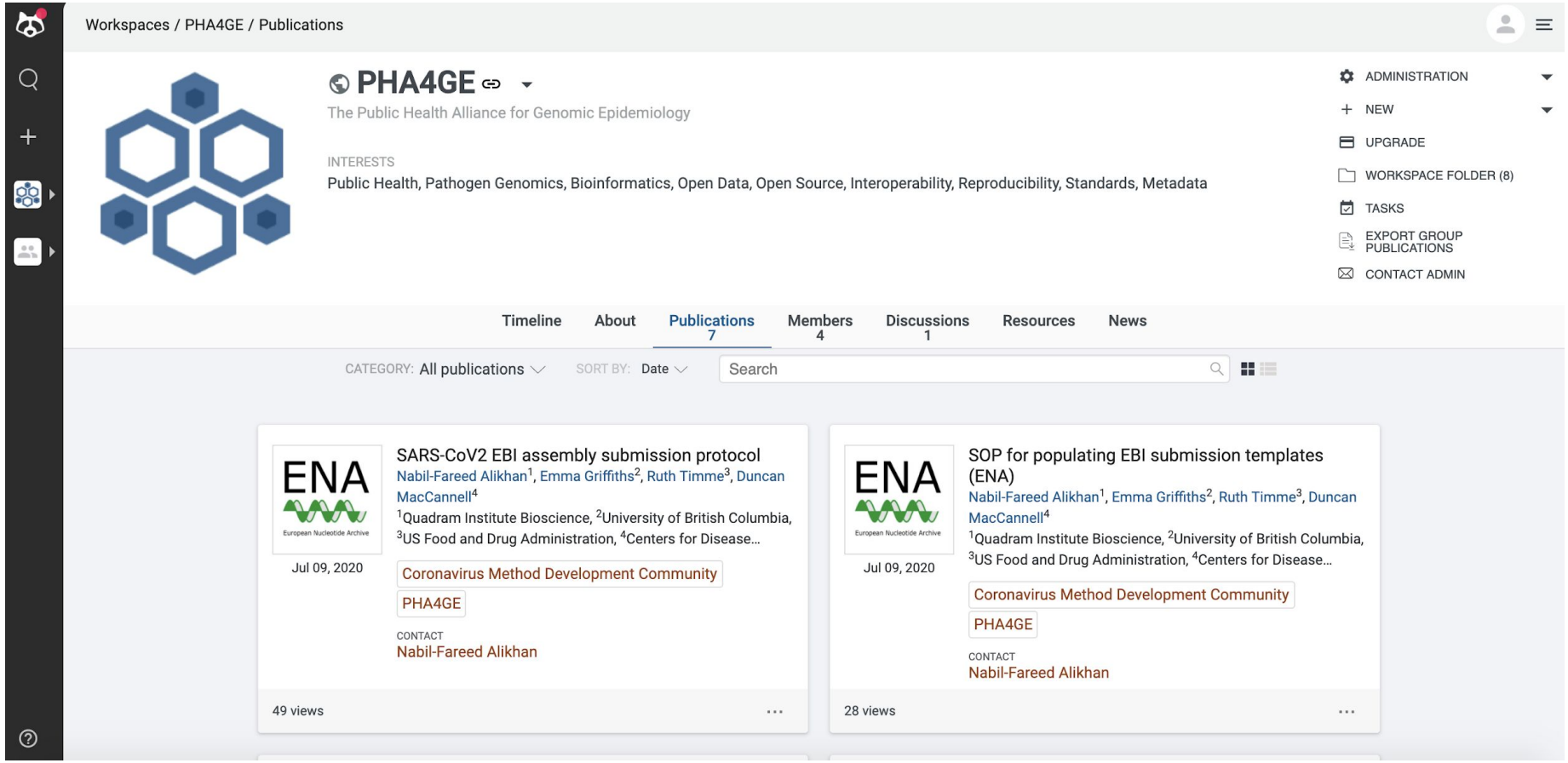
e.g. swab of a hospital bed rail should be recorded:

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

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

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?

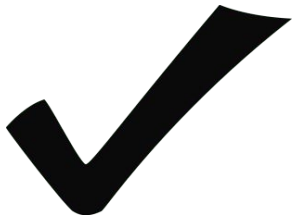
1. Get in touch with us!

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

Public Health Alliance for Genomic Epidemiology

What can the PHA4GE standard do for you?

1. One-stop-shop for consolidating data from different streams

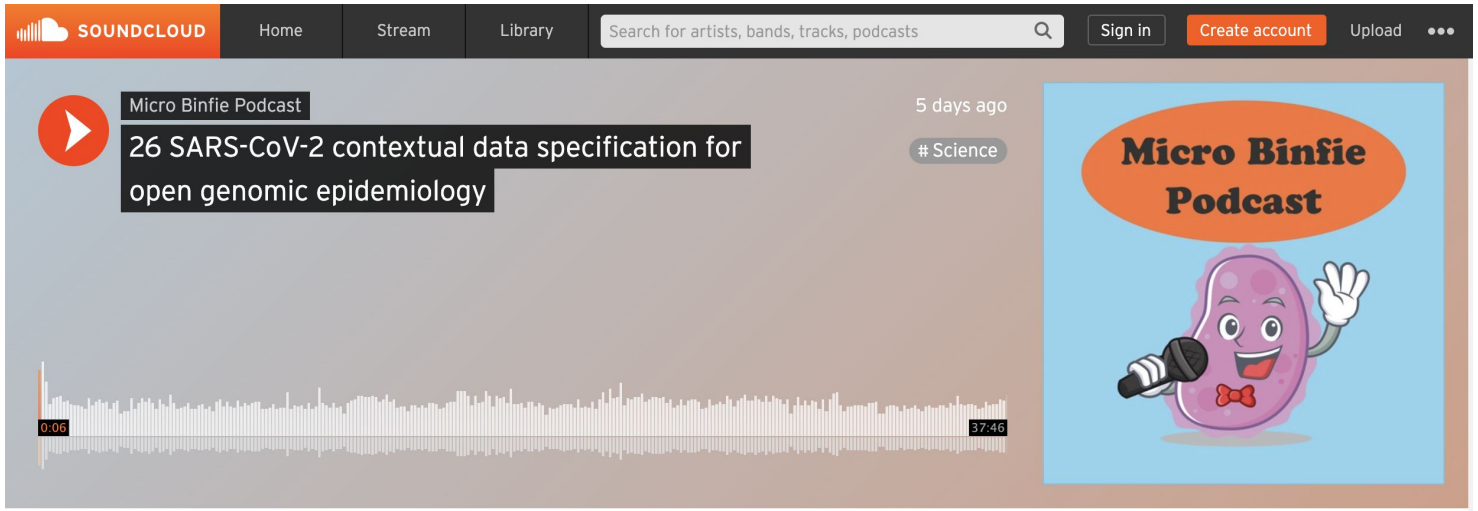


2. Future-proof contextual data

3. Harmonize and integrate data across labs/databases

Public Health Alliance for Genomic Epidemiology

Learn



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

Listen to episode 26 Micro Binfie podcast

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

This work would not be possible without the contributions and dedication of these wonderful people.

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

Public Health Alliance for Genomic Epidemiology