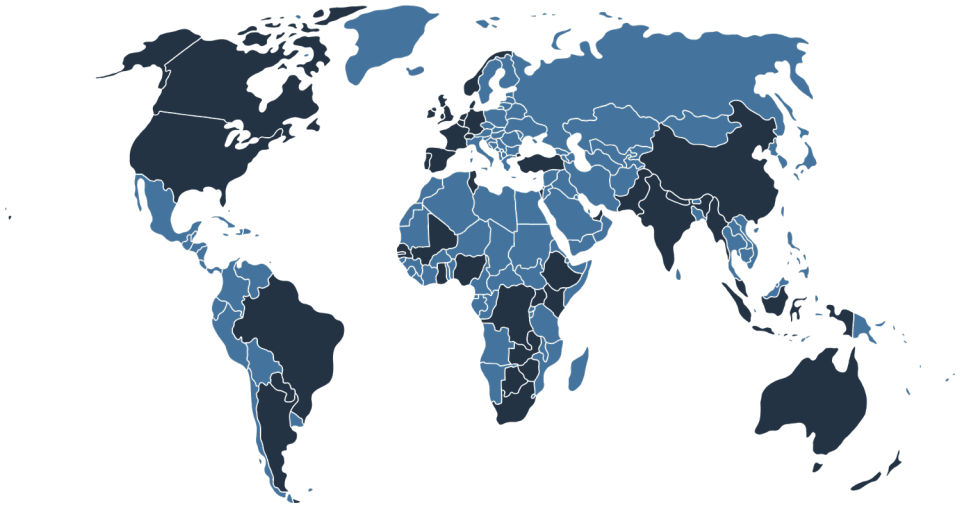
PUBLIC HEALTH ALLIANCE FOR GENOMIC EPIDEMIOLOGY

PHA4GE

Bridging the gap between Public Health & Bioinformatics

Alan Christoffels
South African National Bioinformatics Institute, University of the Western Cape



https://www.pha4ge.org | https://www.github.com/pha4ge | @pha4ge

22-23 July 2024, Washington DC, USA

PUBLIC HEALTH ALLIANCE FOR GENOMIC EPIDEMIOLOGY

Vision

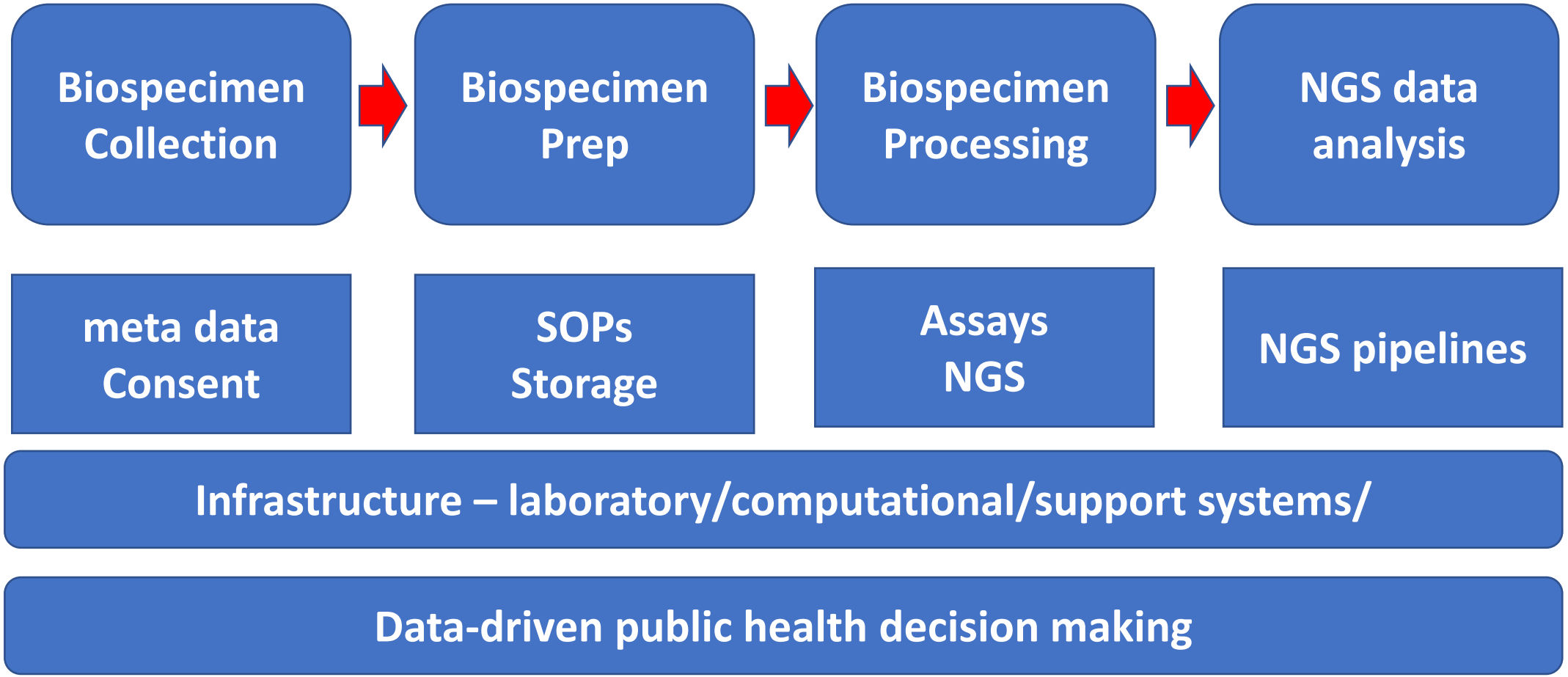
a **modular** and **scalable open-source platform** for public health bioinformatics. Architecture, APIs, ontologies, standards.

Need for a community-driven open model, for public health bioinformatics

Critical incentive gap between academic software development and public health applications

2019 - 2024

PHA4GE embedded in a Public Health setting

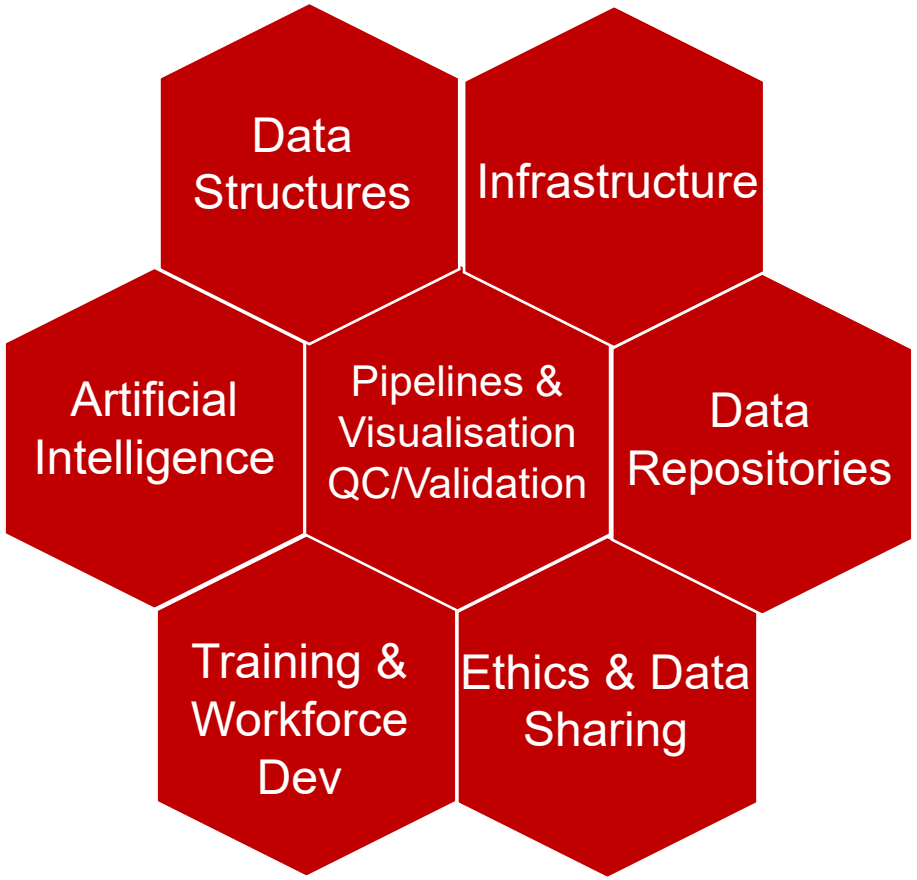


PUBLIC HEALTH ALLIANCE FOR GENOMIC EPIDEMIOLOGY

Strategic Actions

- Sustainability
- Lead best practice implementation for public health laboratories
- Human capital development

Technical Working Groups:



Public Health Alliance for Genomic Epidemiology

Data Structures Working Group

Emma Griffiths

Simon Fraser University

Finlay Maguire

Dalhousie University

https://www.pha4ge.org | https://www.github.com/pha4ge | @pha4ge

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**

Data structure variability in local databases propagates to public Repo

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

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

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

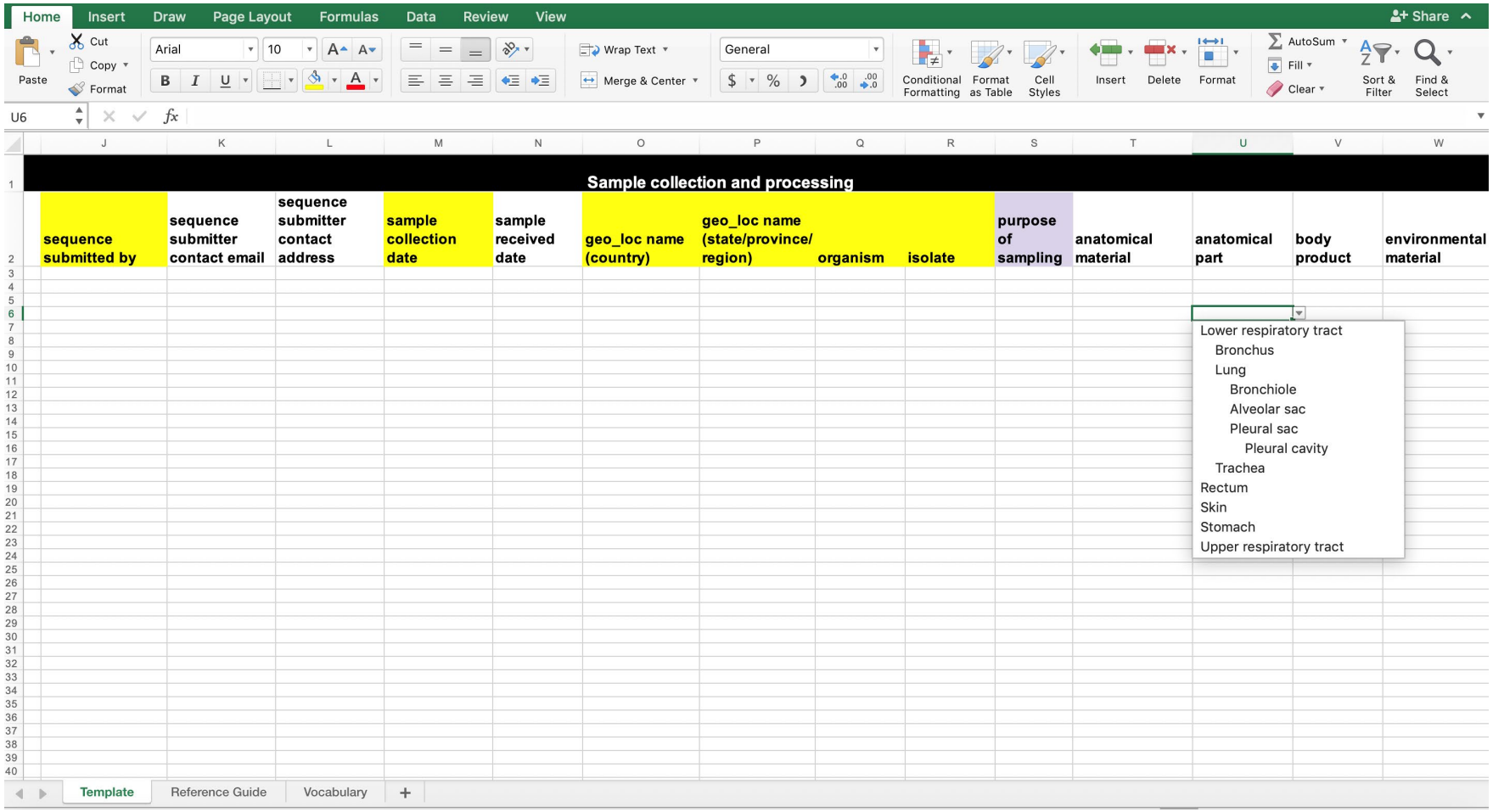
- Case report forms
- Public repository requirements
- Existing metadata standards
- Literature

Mapping to Standards

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

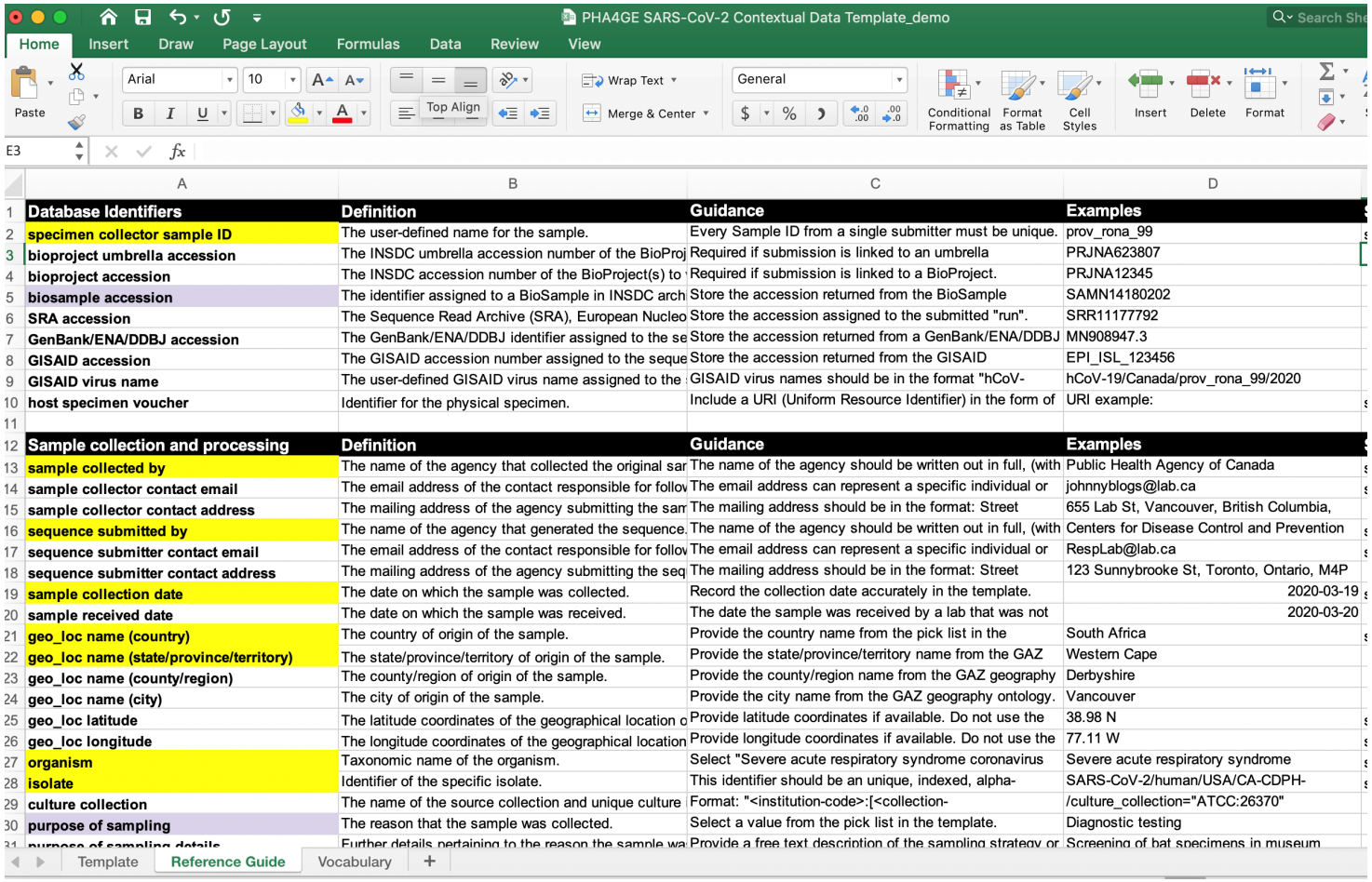
PHA4GE SARS-CoV-2 Specification
https://github.com/pha4ge/SARS-CoV-2-Contextual-Data-Specification

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 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

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:

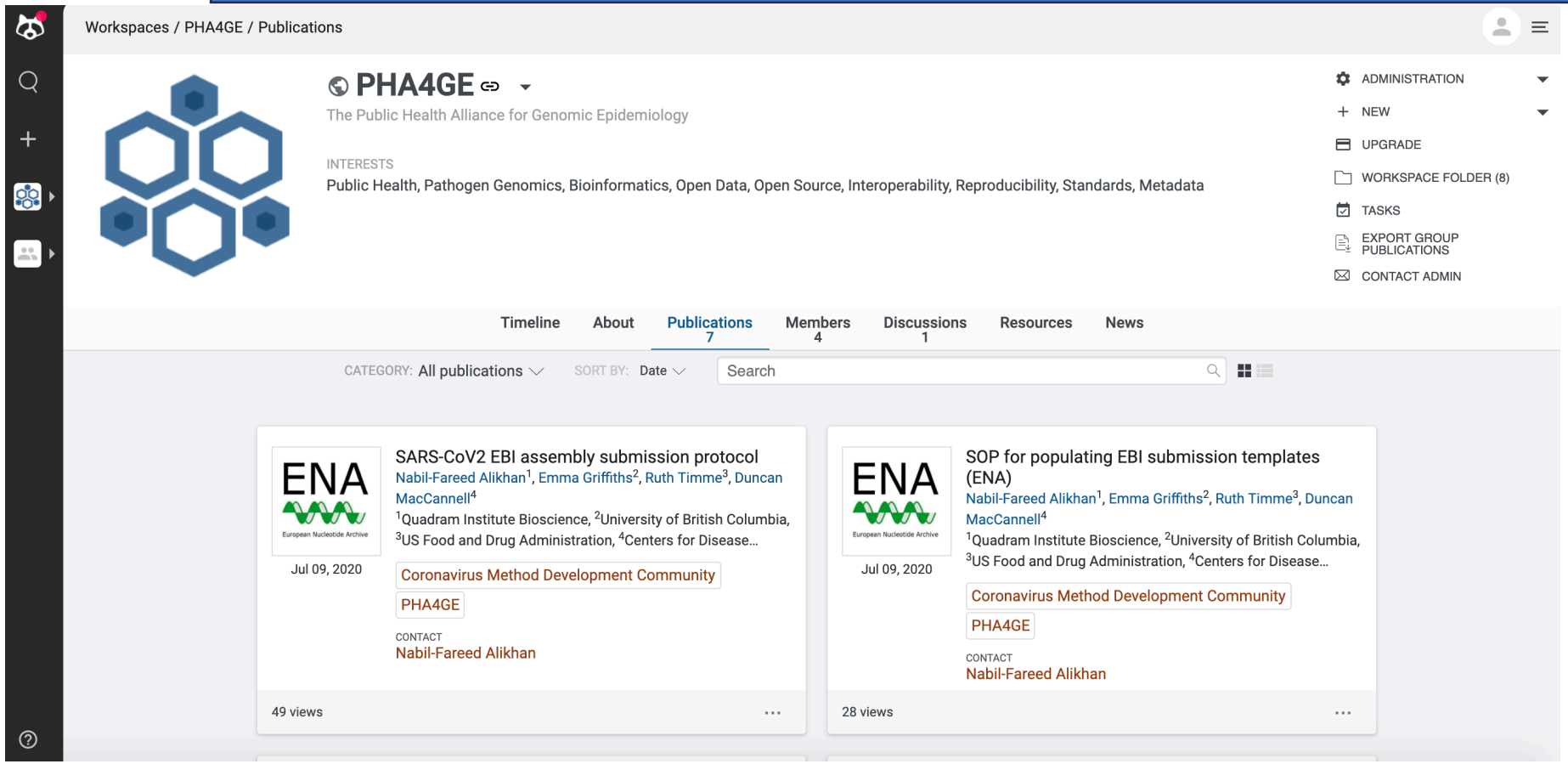
environmental site	environmental material
Sewage Plant	Sewage

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

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

Pipelines & Visualisation Working Group

Victoria Dyster
Mitra Bio

Gültekin Ünal
Ankara University

Emily Smith
Theiagen Genomics

Co-chairs

https://www.pha4ge.org | https://www.github.com/pha4ge | @pha4ge

PUBLIC HEALTH ALLIANCE FOR GENOMIC EPIDEMIOLOGY

Pipelines & Visualisation Working Group

Ten Best Practices for Public Health Bioinformatics Pipelines

PHA4GE Bioinformatics Pipelines & Visualization Working Group
Libuit KG, Guthrie J, Ambrosio F, Kapsak C, Unal Gultekin, Holmes J, Wright S, Nguinkal J, Doughty E, Southgate J, O'Cathail C, Carleton H, Kingwara L, Khan W, Baker K, Diallo A, Connor T, Kanwar S, Maturure P, James S, Cuesta I, Dyster V, Gaskin A, Williams C, Smith E, Rokney A, Petkau A, Varona S, Gnimpieba E, Rey S, Macori G, & Mboowa G

Updates and modifications to this documented are captured in the repository changelog.

- Public accessible Repositories
- Semantic versioning of stable releases
- Workflow management systems
- Packaged software

https://pha4ge.org | https://www.github.com/pha4ge | @pha4ge

PUBLIC HEALTH ALLIANCE FOR GENOMIC EPIDEMIOLOGY

Public Health Implementation

PUBLIC HEALTH ALLIANCE FOR GENOMIC EPIDEMIOLOGY

Thank you

Los Alamos NATIONAL LABORATORY

BIOZENTRUM
University of Basel
The Center for
Molecular Life Sciences

CDC
Centers for Disease Control and Prevention

SANBI
South African National Bioinformatics Institute

University of Birmingham

BROAD INSTITUTE

Doherty Institute

BfR
Bundesinstitut für Risikobewertung

BC Centre for Disease Control

FRED HUTCH

Quadram Institute
Science Health Food Innovation

BIOMÉRIEUX

ASLM
African Society for Laboratory Medicine

H3Africa
Human Heredity & Health in Africa

BILL & MELINDA GATES foundation

LISBOA
UNIVERSIDADE DE LISBOA

wellcome sanger institute

EMBL-EBI

NCBI

Global Alliance for Genomics & Health

CHAN ZUCKERBERG BIOHUB

NATIONAL INSTITUTE FOR COMMUNICABLE DISEASES

IQT
IN·Q·TEL

CHAN ZUCKERBERG INITIATIVE

WRAIR
Walter Reed Army Institute of Research
Soldier Health • World Health

BIG DATA INSTITUTE

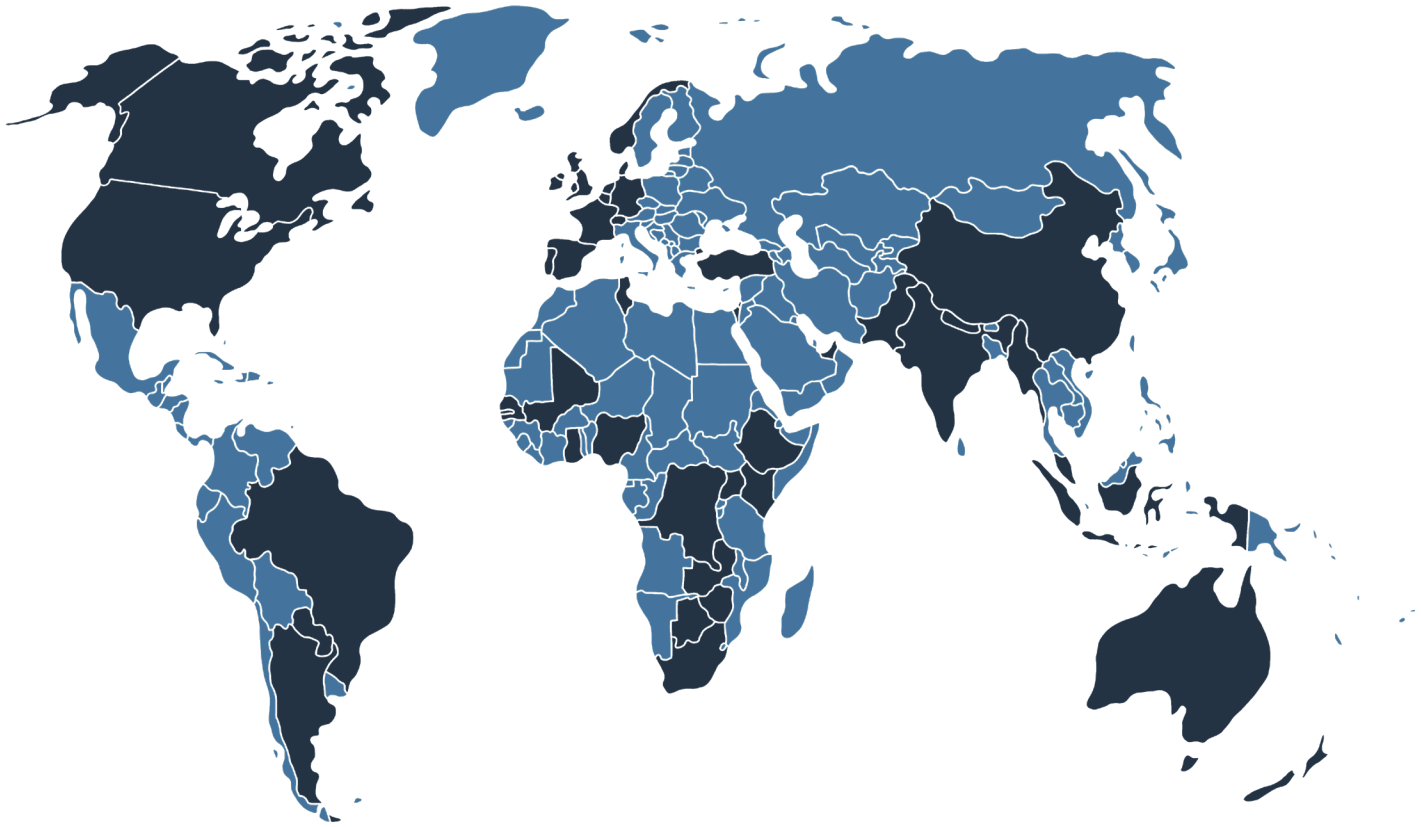
AFRICA CDC
Centres for Disease Control and Prevention
Safeguarding Africa's Health

https://www.pha4ge.org | https://www.github.com/pha4ge | @pha4ge

Extra slides for the audience after the participants

PUBLIC HEALTH ALLIANCE FOR GENOMIC EPIDEMIOLOGY

Sustainability



237 members:
63 – Africa
82 – North America
61 – Europe
17– Asia
8 – Oceania
5 – South America

https://www.pha4ge.org | https://www.github.com/pha4ge | @pha4ge

Public Health Alliance for Genomic Epidemiology

Leading Best Practices

In Progress

Data structures WG

- Data Object Model Paper
- SARS-COV-2 Primer standardization paper
- AMR hAMRonization tool paper

Ethics and Data Sharing WG

- Systematic review paper in progress

Pipelines & Viz WG

- Influenza: Guidance Document in progress covering seasonal and veterinary
- HIV, m.TB, Wastewater Sequencing Guidance Documents in progress.
- Identifying SC2-Recombinants, SC-2 Omicron, MPXV Bioinformatics Solutions
- SC-2 Quality Control Guidance Document

Infrastructure WG

- Framework for Compute infrastructure for Pathogen Genomics Laboratories in progress

Public Health Alliance for Genomic Epidemiology

Leading Best Practices

Published since 2022

- Future-Proofing & maximizing the utility of metadata:…
- https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giac003/6529104

- Framework for the promotion of ethical benefit sharing…
- https://gh.bmj.com/content/7/2/e008096

- **Omicron variant guidance document**
- https://github.com/pha4ge/pipeline-resources/blob/main/docs/omicron-resources.md

- **Identifying SC2 Recombinants Guidance Doc**
- https://github.com/pha4ge/pipeline-resources/blob/main/docs/sc2-recombinants.md

- **SC2 Quality Control Document**
- https://github.com/pha4ge/pipeline-resources/blob/main/docs/qc-solutions.md

- **Proposed Standards for Public Health Bioinformatics Software**
- https://github.com/pha4ge/pipeline-resources/blob/main/docs/pipeline-standards.md

Public Health Alliance for Genomic Epidemiology

Subawards

Focusing on AMR, SARS-CoV-2 and Ethics & Data Sharing themes

Botswana
Ethiopia
Kenya
Malawi
Nigeria
South Africa
Sudan
The Gambia
Uganda
Zimbabwe

Cambodia
Congo
DR Congo
Fiji
Malaysia
Nepal
Pakistan
Philippines
Zambia

PUBLIC HEALTH ALLIANCE FOR GENOMIC EPIDEMIOLOGY

Ethics and Data sharing Working Group

Nicki Tiffin (UWC, SA)

Framework for sharing data during outbreaks
https://ethics-forum.pha4ge.org/

ELSI subgrants

Democratic Republic of Congo & Gabon:
(Ethics Training)
Setup, run dedicated online ethics training resources

Kenya:
(Ethics Implementation)
Engage with ethics review boards - review of genomic research applications

Zimbabwe:
(Scientific Citizenship)
Videos made by local researchers targeting school learners using local language

Nigeria:
(Community engagement)
Vaccine Hesitancy in Nigeria