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

Preprint · August 2020

DOI: 10.20944/preprints202008.0220.v1

The PHA4GE SARS-CoV-2 contextual data specification for open genomic epidemiology

Authors

Emma J Griffiths [1*], Ruth E Timme [2], Andrew J Page [3], Nabil-Fareed Alikhan [3], Dan Fornika [4], Finlay Maguire [5], Catarina Inês Mendes [6], Simon H Tausch [7], Allison Black [8], Thomas R Connor [9,10], Gregory H Tyson [11], David M Aanensen [12, 13], Brian Alcock [14], Josefina Campos [15], Alan Christoffels [16], Anders Gonçalves da Silva [17], Emma Hodcroft [18], William WL Hsiao [1, 19, 20], Lee S Katz [21,], Samuel M Nicholls [22], Paul E Oluniyi [23, 24], Idowu B Olawoye [23, 24], Amogelang R Raphenya [14], Ana Tereza R Vasconcelos [25], Adam A Witney [26], and Duncan R MacCannell [27], on behalf of the Public Health Alliance for Genomic Epidemiology (PHA4GE) consortium.

Affiliations

1 Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, British Columbia, Canada
2 Center for Food Safety and Applied Nutrition, U.S. Food and Drug Administration, College Park, MD, USA.
3 Quadram Institute Bioscience, Norwich, Norfolk, UK
4 BC Centre for Disease Control Public Health Laboratory, Vancouver, Canada
5 Faculty of Computer Science, Dalhousie University, Halifax, Canada
6 Instituto de Microbiologia, Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, Lisboa, Portugal
7 Department of Biological Safety, German Federal Institute for Risk Assessment, Berlin, Germany
8 Department of Epidemiology, University of Washington, Washington, USA.
9 Organisms and Environment Division, School of Biosciences, Cardiff University, Cardiff, Wales, UK.
10. Public Health Wales, University Hospital of Wales, Cardiff, UK
11 Center for Veterinary Medicine, U.S. Food and Drug Administration, Laurel, Maryland, USA
12 Centre for Genomic Pathogen Surveillance, Wellcome Genome Campus, Cambridge, UK
13 The Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Medicine, University of Oxford, Oxford, UK.
14 Department of Biochemistry and Biomedical Sciences and the Michael G. DeGroote Institute for Infectious Disease Research, McMaster University, Hamilton, Ontario, Canada
15 INEI-ANLIS "Dr Carlos G. Malbrán", Buenos Aires, Argentina
16 South African Medical Research Council Bioinformatics Unit, South African National Bioinformatics Institute, University of the Western Cape, Bellville, South Africa
17 Microbiological Diagnostic Unit Public Health Laboratory, The Peter Doherty Institute for Infection and Immunity, The University of Melbourne, Melbourne, Victoria, Australia

18 Biozentrum, University of Basel, Basel, Switzerland & Swiss Institute of Bioinformatics, Lausanne, Switzerland
19 British Columbia Centre for Disease Control Public Health Laboratory, Vancouver, British Columbia, Canada
20 Department of Health Sciences, Simon Fraser University, Burnaby, British Columbia, Canada
21 Center for Food Safety, University of Georgia, Georgia, USA
22 University of Birmingham, Birmingham, UK
23 African Center of Excellence for Genomics of Infectious Diseases (ACEGID), Redeemer's University, Ede, Osun State, Nigeria
24 Department of Biological Sciences, College of Natural Sciences, Redeemer's University, Ede, Osun State, Nigeria
25 Bioinformatics Laboratory National Laboratory of Scientific Computation LNCC/MCTI, Rio de Janeiro, Brazil
26 Institute for Infection and Immunity, St George's, University of London, London, UK
27 National Center for Emerging and Zoonotic Infectious Diseases, Centers for Disease Control and Prevention, Georgia, USA

*Corresponding author: Emma.Griffiths@bccdc.ca

Keywords: genomics, metadata, SARS-CoV-2, bioinformatics, data standards

Abstract

The Public Health Alliance for Genomic Epidemiology (PHA4GE) (https://pha4ge.org) is a global coalition that is actively working to establish consensus standards, document and share best practices, improve the availability of critical bioinformatic tools and resources, and advocate for greater openness, interoperability, accessibility and reproducibility in public health microbial bioinformatics. In the face of the current pandemic, PHA4GE has identified a clear and present need for a fit-for-purpose, open source SARS-CoV-2 contextual data standard. As such, we have developed an extension to the INSDC pathogen package, providing a SARS-CoV-2 contextual data specification based on harmonisable, publicly available, community standards. The specification is implementable via a collection template, as well as an array of protocols and tools to support the harmonisation and submission of sequence data and contextual information to public repositories. Well-structured, rich contextual data adds value, promotes reuse, and enables aggregation and integration of disparate data sets. Adoption of the proposed standard and practices will better enable interoperability between datasets and systems, improve the consistency and utility of generated data, and ultimately facilitate novel insights and discoveries in SARS-CoV-2 and COVID-19.

The importance of contextual data for interpreting SARS-CoV-2 sequences

The SARS-CoV-2 pandemic has been referred to as a once-in-a-century event (1). Beginning in late 2019 in Wuhan, China, the virus has now spread to virtually every country and territory in the world, causing hundreds of thousands of deaths and millions of confirmed cases of COVID-19 (2,3). Understanding, monitoring and preventing transmission have been primary goals of the public health response to SARS-CoV-2.

Tracking the spread and evolution of the virus at global, national and local scales has been aided by the analysis of viral genome sequence data alongside SARS-CoV-2 epidemiology. Large scale sequencing efforts are often formalised as consortia across the world, including the COG-UK in the UK (4), SPHERES in the USA (5), CanCOGeN in Canada (6), Latin American Genomics SARS-CoV-2 Network (7, 8), 2019nCoVR in China (9), and the South Africa NGS Genomic Surveillance Network (10). These combined efforts will result in the generation of hundreds of thousands of genome sequences within the first year of the pandemic. Deposition of these sequences into public repositories such as the Global Initiative on Sharing All Influenza Data (GISAID) (11) and the International Nucleotide Sequence Database Collaboration (INSDC)(12) has enabled rapid global sharing of data. At the time of writing, 97 countries had undertaken open sequencing initiatives (GISAID accessed 2020-07-02) generating over 58,545 sequences which are being reused and analysed on a massive scale. The open data sharing paradigm has had tremendous success in the genomic epidemiology of foodborne pathogens (13, 14), and has the potential to reveal a deeper understanding of SARS-CoV-2 origin, pathogenicity, and basic biology when submissions from its wild hosts are included alongside human sample (15). The open sharing of SARS-CoV-2 data has already paid dividends for diagnostics and catalyzed a number of vaccine initiatives (16, 17). Mutations in genomes rendering assay probes less sensitive or ineffective is highly problematic in a pandemic where testing is a crucial aspect of infection control. Global monitoring of mutations in platforms like Nextstrain and CoV-Glue-UK have better enabled agility and confidence in the diagnostic domain (18).

Public health sequence data is of limited value without contextual data, which consists of laboratory (e.g. date and location of testing, cycle threshold (CT) values), clinical (e.g. hospitalization, outcomes), epidemiological (e.g. age, gender, exposures) and methods (sampling, sequencing, bioinformatics) information. For example, phylodynamics, the combined analysis of epidemiological, immunological, and evolutionary characteristics, is predicated on having accurate sampling time and location data for each genome which aid public health practitioners in understanding the spatiotemporal patterns of disease transmission (19-21). Additionally, contextual data may be used to determine whether specific lineages are circulating in specific settings e.g. long-term care facilities (22), meat packing plants (23), conferences (24) or other public gatherings (25), or travel-related (26, 27). The importance of the contextual data in evaluating the epidemiological relevance of genomic relationships (28) is particularly important in low-diversity pathogens, such as SARS-CoV-2. Genomic variations are a key source of information that can help public health researchers better understand putative changes in transmission, virulence, epidemiology and therapeutics of an emerging pathogen. Evaluating which variants represent real, circulating viruses, as opposed to artifacts of sample handling or sequencing depends on the capture of methodological information, such as experimental design, laboratory procedures, bioinformatic processing, and quality control metrics, in order to understand the context and limitations of analyses e.g. detecting systematic batch effect errors related to certain sequencing centres and methods (29-31). These are just a few examples, and there are many additional ways to interpret public health genomic data to inform decision making for public health responses and develop greater scientific understanding of the pathogen.

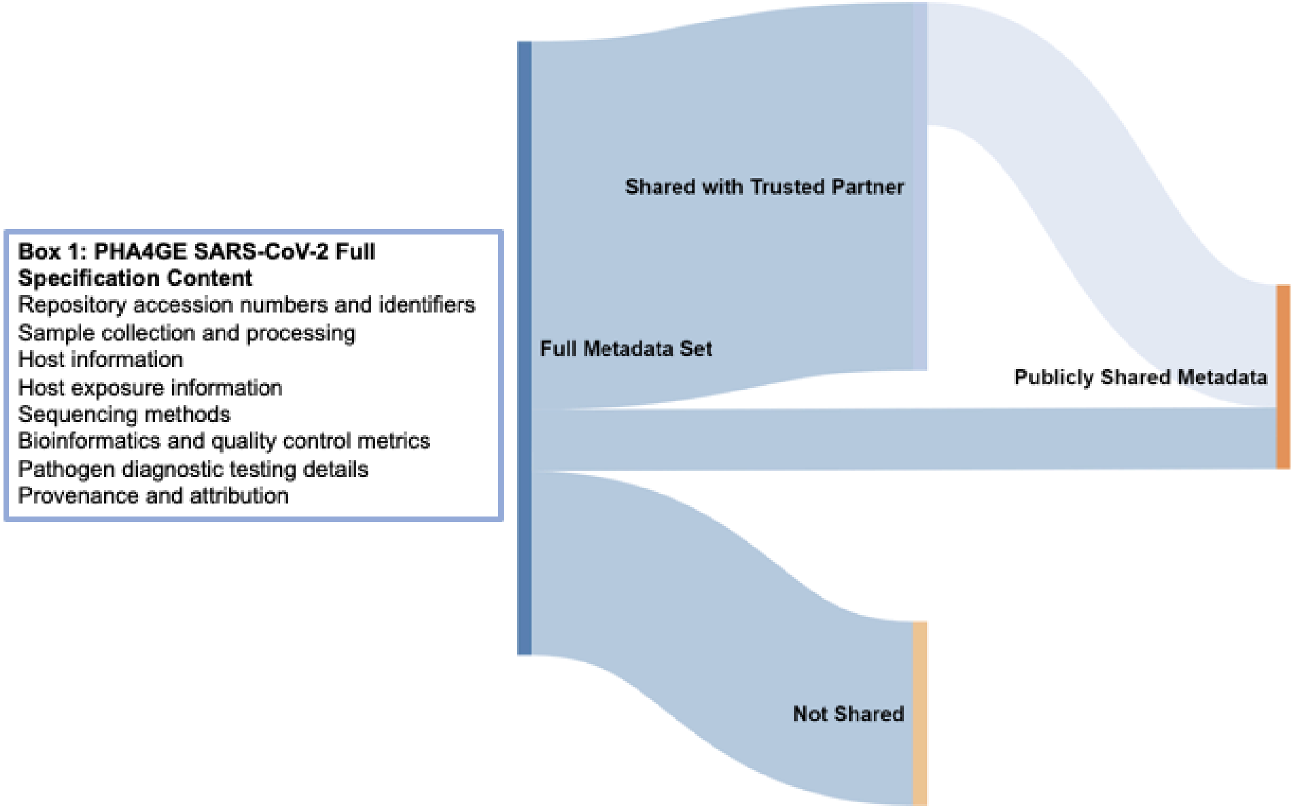
Contextual data that are structured and consistent, particularly complying with community standards like minimum information checklists (MIxS (32), MIGS (33), Sample Application Standard (34)) and ontologies (OBO Foundry (35)), are easier to understand and process, and can be more easily aggregated and reused for different types of analyses. However, contextual data is often collected on a project-specific basis according to local needs and reporting requirements, and is often structured according to organization or initiative-specific data dictionaries. Furthermore, attribute packages and metadata standards developed by different organizations are scoped to cover as many use cases and pathogens as possible, and so can include fields of information not applicable to SARS-CoV-2 or that may be subject to privacy concerns, or exclude fields commonly used in public health surveillance and investigations. As different types of contextual data are subject to different ethical, practical and privacy concerns, not all components of existing standards are immediately or widely shareable. As a result, the range of generic metadata standards being applied to SARS-CoV-2 data presents challenges for data harmonization (36) and analysis critical for fighting the disease and ending the pandemic.

While the examples here focus on public health surveillance, we must recognize that good data management (tracking and documenting) goes beyond data sharing. Good data stewardship practices are not only critical for auditability and reproducibility, but for posterity - documenting critical information about samples, methods, risk factors and outcomes etc, can help build a roadmap for dealing with future public health crises.

In light of these challenges, PHA4GE has identified a clear and present need for a fit-for-purpose, open source SARS-CoV-2 contextual data specification which can be used to consistently structure information as part of good data management practices and for data sharing with trusted partners and/or public repositories. The specification was developed by consensus among domain experts, and incorporates existing community standards in light of SARS-CoV-2 public health needs in order to ensure privacy while maximizing information linkage, content and interoperability across datasets and databases, to better enable analyses to fight COVID-19.

SARS-CoV-2 Contextual Data Specification: The Framework

The purpose of the PHA4GE SARS-CoV-2 specification is to provide a mechanism for consistent structure, collection and formatting of fields and values containing SARS-CoV-2 contextual data. We emphasize that the purpose of this specification is not to force data sharing, but rather to provide a framework to structure data consistently across disparate laboratory and epidemiological databases so that they can be harmonized for different uses (Figure 1). Data sharing is just one use case and can involve sharing between divisions within a single agency, sharing between partners based on memorandums of understanding, or submission to public repositories.



Box 1: PHA4GE SARS-CoV-2 Full Specification Content
Repository accession numbers and identifiers
Sample collection and processing
Host information
Host exposure information
Sequencing methods
Bioinformatics and quality control metrics
Pathogen diagnostic testing details
Provenance and attribution

Figure 1: Contextual data flow.
Contextual data can be captured and structured using the PHA4GE specification so that it can be more easily harmonized across different data sources and providers. Different subsets of the harmonized data

can be 1) shared with public repositories e.g. GISAID and INSDC, 2) shared with trusted partners e.g. national sequencing consortia, public health partners, and 3) kept private and retained locally with the potential for sharing in the future for particular surveillance or research activities. While fields have been colour-coded in the template to indicate whether they are considered “required”, “strongly recommended” and “optional”, how the specification is implemented, and how, if any, of the data is shared, is ultimately at the discretion of the user. Box 1 describes the information types covered in the full specification.

The PHA4GE SARS-CoV-2 contextual data specification was created through broad consultation with representatives from public health laboratories, research institutes and universities in eight countries (Canada, Australia, Germany, Portugal, South Africa, Switzerland, the United Kingdom, the United States of America) who are involved with the SARS-CoV-2 genome sequencing and analysis efforts at various levels. Based on this consultation and consensus, the specification contains different fields covering a wide array of data types described in Box 1 (Figure 1). The specification attempts to harmonize different data standards (INSDC, GISAID, MIxS, MIGS, Sample Application Standard) by reusing fields or mapping to fields, as much as possible. As PHA4GE embraces FAIR data stewardship principles (Findability, Accessibility, Interoperability and Reuse of digital assets), we strived to implement FAIR principles in the design and implementation of the specification for data management and data sharing. At their core, these principles emphasize machine-actionability and consistency of data, and are critical for dealing with the volume and complexity of genomic sequence and contextual data.

The versioned specification is available as a contextual data collection template (.xlsx) and in machine-amenable JSON format from GitHub (https://github.com/pha4ge/SARS-CoV-2-Contextual-Data-Specification; version at time of publication https://zenodo.org/record/3947048#.Xxs7gvhKg_U). The collection template also offers ontology-based standardized terms for a number of fields in the form of pick lists. The template is also supported by a number of materials such as a Reference Guide, which provides definitions and field-level guidance, as well as examples of how data might appear when structured according to the specification. A Standard Operating Procedure (SOP), which contains instructions for using the collection template has also been provided. Mapping of fields to standards and public repository submission requirements, and links to protocols that have been developed by PHA4GE for SARS-CoV-2 sequence submission have also been provided. A table outlining the different materials can be found in Table 1.

Table 1: Resources that form the PHA4GE SARS-CoV-2 contextual data specification package available from https://github.com/pha4ge/SARS-CoV-2-Contextual-Data-Specification (version at time of publication https://zenodo.org/record/3947048#.Xxs7gvhKg_U)

Resource[1]	**Description**
Collection template and controlled vocabulary pick lists	Spreadsheet-based collection form containing different fields (identifiers and accessions, sample collection and processing, sequencing, host information, host exposure information, bioinformatics and QC metrics, author acknowledgements). Fields are colour-coded to indicate required, recommended or optional status. Many fields offer pick lists of controlled vocabulary. Vocabulary lists are also available in a separate tab.
Reference guide	Field definitions, guidance, and examples are provided as a separate tab in the collection template .xlsx file.
Collection template SOP	Step-by-step instructions for using the collection template are provided in the SOP. Ethical, practical, and privacy considerations are also discussed. Examples and instructions for structuring sample descriptions as well as sourcing additional standardized terms (outside those provided in pick lists) are also discussed.
PHA4GE fields to metadata standards mapping	PHA4GE fields are mapped to existing metadata standards such as the Sample Application Standard, MIxS 5.0, and the MIGS Virus Host-associated attribute package. Mappings are available in the Reference guide tab. Mappings highlight which fields of these standards are considered useful for SARS-CoV-2 public health surveillance and investigations, and which fields are considered not applicable.
EMBL-EBI, NCBI and GISAID submission requirements to PHA4GE field mappings	Many PHA4GE fields have been sourced from public repository submission requirements. The different repositories have different requirements and field names. Repository submission fields have been mapped to PHA4GE fields to demonstrate equivalencies and divergences.
Data submission protocol (NCBI)	The SARS-CoV-2 submission protocol for NCBI provides step-by-step instructions and recommendations aimed at improving interoperability and consistency of submitted data.
Data submission protocol (EMBL-EBI)	The SARS-CoV-2 submission protocol for ENA provides step-by-step instructions and recommendations aimed at improving interoperability and consistency of submitted data.
Data submission protocol (GISAID)	The SARS-CoV-2 submission protocol for GISAID provides step-by-step instructions and recommendations aimed at improving interoperability and consistency of submitted data.
JSON structure of PHA4GE specification	A JSON structure of the PHA4GE specification has been provided for easier integration into software applications.

[1] There are a number of resources that form the PHA4GE SARS-CoV-2 contextual data specification package which are described in the table. The package has been compiled to support user implementation and data sharing, with integration into workflows and new software applications in mind.

Getting Started - How To Use The Standard

In designing the specification we first began with considering the goals of data collection and harmonization. Consulted partners felt that the primary priority of standardizing data should be improved support for SARS-CoV-2 genomic surveillance activities and the submission of sequence data and minimal metadata to public repositories. The two most important attributes for tracking transmission from pathogen genomic data are temporal information describing when a sample was collected and spatial information describing where a virus was sampled. Comparisons of minimal contextual data requirements across different national sequencing efforts, as well as submission requirements for INSDC and GISAID databases, yielded a minimal set of 10 fields which we have annotated as "required" in the specification (colour-coded yellow in the collection template, see Table 1). Those fields, their definitions, and guidance notes are described in Table 2. A number of other fields have been annotated as "strongly recommended" (colour-coded purple in the collection template) for capturing sample collection and processing methods, critical epidemiological information about the host, and acknowledging scientific contributions. Fields colour-coded white are considered optional.

Table 2: Minimal (required) contextual data fields

Field Name[1]	**Definition**	**Guidance**
specimen collector sample ID	The user-defined name for the sample.	Every Sample ID from a single submitter must be unique. It can have any format, but we suggest that you make it concise, unique and consistent within your lab, and as informative as possible.
sample collected by	The name of the agency that collected the original sample.	The name of the agency should be written out in full, (with minor exceptions) and consistent across multiple submissions.
sequence submitted by	The name of the agency that generated the sequence.	The name of the agency should be written out in full, (with minor exceptions) and be consistent across multiple submissions.
sample collection date	The date on which the sample was collected.	Record the collection date accurately in the template. Required granularity includes year, month and day. Before sharing this data, ensure this date is not considered identifiable information. If this date is considered identifiable, it is acceptable to add "jitter" to the collection date by adding or subtracting calendar days. Do not change the collection date in your original

		records. Alternatively, ”received date” may be used as a substitute in the data you share. The date should be provided in ISO 8601 standard format "YYYY-MM-DD".
geo_loc name (country)	Country of origin of the sample.	Provide the country name from the pick list in the template.
geo_loc name (state/province/region)	State/province/region of origin of the sample.	Provide the state/province/region name from the GAZ geography ontology. Search for geography terms here: https://www.ebi.ac.uk/ols/ontologies/gaz
organism	Taxonomic name of the organism.	Use “Severe acute respiratory syndrome coronavirus 2”
isolate	Identifier of the specific isolate.	This identifier should be an unique, indexed, alpha-numeric ID within your laboratory. If submitted to the INSDC, the "isolate" name is propagated throughout different databases. As such, structure the "isolate" name to be ICTV/INSDC compliant in the following format: "SARS-CoV-2/host/country/sampleID/date"
host (scientific name)	The taxonomic, or scientific name of the host.	Common name or scientific name are required if there was a host. Scientific name examples e.g. Homo sapiens. Select a value from the pick list. If the sample was environmental, put "not applicable".
host disease	The name of the disease experienced by the host.	This field is only required if there was a host. If the host was a human select COVID-19 from the pick list. If the host was asymptomatic, this can be recorded under “host health state details”. "COVID-19" should still be provided if the patient is asymptomatic. If the host is not human, and the disease state is not known or the host appears healthy, put “not applicable”.

[1]Through consultation and consensus, ten fields were prioritized for SARS-CoV-2 surveillance, which are considered required in the specification. Field names, definitions, and guidance are presented.

As many contextual data types are stored in different locations and databases (e.g. LIMS, epidemiology case report forms and databases), a benefit of implementing the PHA4GE collection template is that it enables the capture of these different pieces of information in one place. The collection template also offers picklists for a variety of fields e.g. a curated INSDC country list for “geo_loc name (country)”, the standardised name of the virus under the “organism” field (i.e. Severe acute respiratory coronavirus 2), and a multitude of standardised terms for cell lines in the “lab host” field. The picklists provided are neither exhaustive, nor comprehensive, but have been curated from current literature representing active sampling and surveillance activities. If a pick list is missing standardised terms of interest, the reference guide also provides links to different ontology look-up services enabling users to identify additional standardized terms. The reference guide provides definitions for the fields, additional guidance regarding the structure of the values in the field, and any suggestions for addressing issues pertaining to privacy and identifiability. The template SOP provides users with step-by-step instructions for populating the template, looking up standardized terms, and how best to structure sample descriptions. The SOP also highlights a number of ethical, practical, and privacy considerations for data sharing.

Implementation of the PHA4GE specification around the world

How, and how much of, the specification is implemented is ultimately at the discretion of the user. To date, versions of the specification are being implemented in the CanCOGeN (Canada) and SPHERES (USA) SARS-CoV-2 sequencing initiatives, the AusTrakka (Australia) national data sharing platform (37), by the Global Emerging Pathogens Treatment Consortium (Africa) (38), and in the Baobab LIMS (39) at the South African National Bioinformatics Institute (SANBI) (40).

Submitting Data to Public Sequence Repositories

For a large global genomic surveillance program to be successful, each new entry (genome, contextual data (usually referred to as metadata), plus raw reads) must be made publicly available and properly linked in one of the collaborating databases. Most existing SARS-CoV-2 sequences have only been deposited in GISAID, with a small proportion of submitters (~20%, 2020-06-04) also depositing matching raw read data in the INSDC (i.e. National Center for Biotechnology Information (NCBI), European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI) and DNA Data Bank of Japan (DDBJ)).

Within the INSDC, the metadata describing the samples are stored as accessioned BioSamples (41) with a consistent set of attribute names and standardized values. BioSamples add value, promote reuse, and enable interoperability of data submitted from laboratories that may only be connected by following the same metadata standard. The INSDC databases provide a generic pathogen metadata template for the BioSample that is heavily utilized for bacterial genomic surveillance (42) and extended for particular use cases (32). GISAID uses a different format and data structure for associating metadata primarily for influenza surveillance and now extended to include SARS-CoV-2. The ENA provides a virus metadata checklist (ENA virus pathogen reporting standard checklist) developed as part of the COMPARE project (43), which is very similar to the GISAID submission requirements. Building off of these existing standards, we developed a metadata specification for SARS-CoV-2 genomic surveillance that is broad enough for internal laboratory use while providing formated submission templates for public release to INSDC and GISAID. The detailed mapping of PHA4GE fields to public repository submission requirements as well as guidance and advice are available as supporting documents (see Table 1). We have also provided detailed protocols for data submission to the three participating repositories, GenBank/SRA (NCBI), ENA (EMBL-EBI), and GISAID. An overview of how the PHA4GE specification is integrated into public repository submissions is presented in Figure 2.

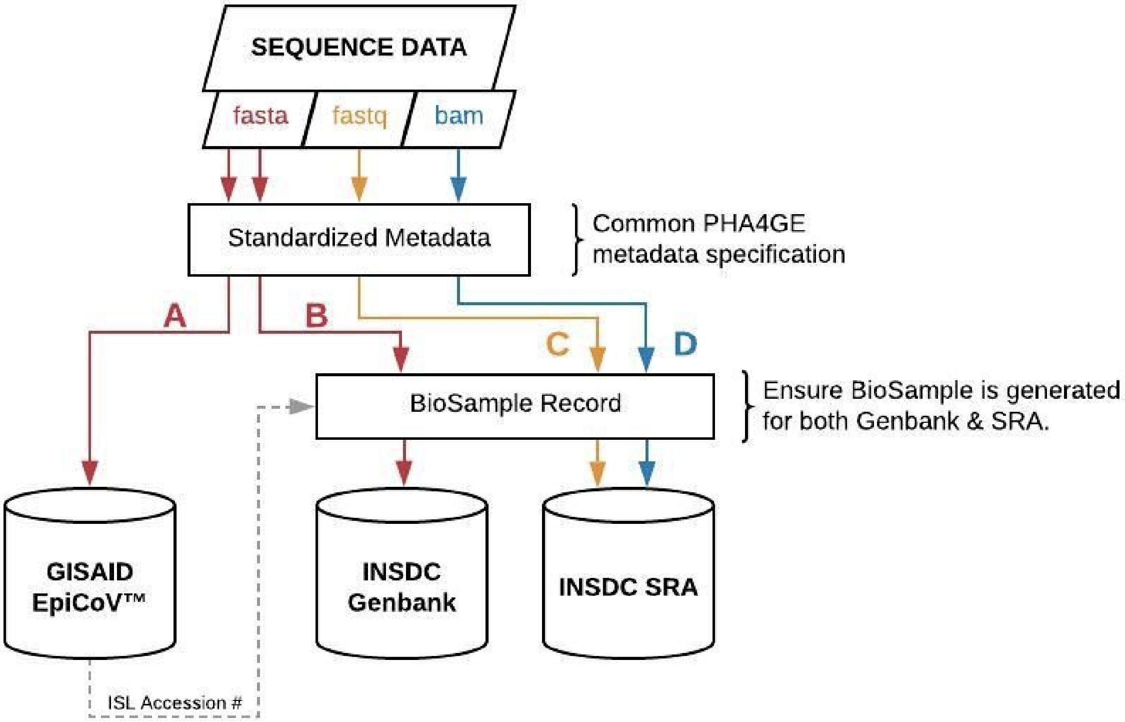


Figure 2: Overview of how the PHA4GE SARS-CoV-2 contextual data specification can be integrated into public repository submission.
The PHA4GE collection template provides a one-stop-shop for different data types that are important for global surveillance. The protocols provided as part of the specification package describe how PHA4GE fields can be mapped to different repository submission forms. Consensus sequences (FASTA), accompanied by a subset of PHAGE fields, can be submitted to the GISAID EpiCoV database (A). Consensus sequences (FASTA) (B) as well as raw/processed data (FASTQ, BAM) (C, D) can be submitted to INSDC databases (e.g. GenBank, SRA) with different subsets of PHA4GE fields as part of a BioSample record. BioSamples are propagated throughout INSDC databases.

PHA4GE recommendations for FAIR SARS-CoV-2 INSDC data submissions are as follows:

1. submit raw sequencing data and assembled/consensus genomes to INSDC and GISAID
2. create a BioSample record when submitting using the PHA4GE guidance, populating the mandatory and recommended fields where possible
3. curate your public records (sequence data and BioSample), updating them when subsequent information becomes available or retracting if/when records become untrustworthy.

Conclusion

The collective response to the SARS-CoV-2 pandemic has resulted in an unprecedented deployment of genomic surveillance worldwide, bringing together public health agencies, academic research institutions, and industry partners. This unified action provides opportunities to more effectively understand and respond to the pandemic. Yet it also provides an enormous challenge, as realizing the full potential of this opportunity will require standardization and harmonization of data collection across these partners. As countries around the world face exponential growth in the number of COVID-19 cases, and prepare for new waves of infections throughout the pandemic, a unique opportunity for harmonization in data collection exists. With our SARS-CoV-2 metadata specification we have endeavored to create a mechanism for promoting consistent, standardized contextual data collection that can be applied broadly. We hope that, given sufficient uptake, this specification will improve the consistency of collected data, making them reusable by agencies as they continue working towards an increased understanding of SARS-CoV-2 epidemiology and biology, and harmonizing them such that community-based data sharing efforts are not excessively burdened.

Furthermore, the framework for SARS-CoV-2 presented in this work can also be used to build a roadmap for dealing with future public health crises.

Acknowledgements

We gratefully acknowledge Alecia Naidu and Jamie Southgate of the PHA4GE secretariat for their support of this work. We also thank all those who provided feedback during consultations, particularly Anthony Underwood, Peter van Heusden, Torsten Seemann, Danny Park, Andrew Rambaut and Zamin Iqbal.

Disclaimer

The views expressed in this article are those of the authors and do not necessarily reflect the official policy of the U.S. Department of Health and Human Services, the U.S. Food and Drug Administration, the U.S. Centers for Disease Control and Prevention, or the U.S. Government.

Funding statements

All authors have read this manuscript and consented to its publication. We wish to thank the Bill & Melinda Gates Foundation for supporting the establishment and work of the PHA4GE consortium. AJP and NFA gratefully acknowledge the support of the Biotechnology and Biological Sciences Research Council (BBSRC); and were supported by the Quadram Institute Bioscience BBSRC funded Core Capability Grant (project number BB/CCG1860/1) and the BBSRC Institute Strategic Programme Microbes in the Food Chain BB/R012504/1 and its constituent project BBS/E/F/000PR10352. FM was supported by a Donald Hill Family Fellowship in Computer Science. CIM was supported by the Fundação para a Ciência e Tecnologia (grant SFRH/BD/129483/2017). Work by EJG and WWLH was funded by a Genome Canada Bioinformatics and Computational Biology 2017 Grant #286GET.

References

1. Gates B. Responding to Covid-19 - A Once-in-a-Century Pandemic? N Engl J Med. 2020 Apr 30;382(18):1677–9.
2. WHO. Coronavirus disease (COVID-19) pandemic [Internet]. [cited 2020 Jun 9]. Available from: https://www.who.int/emergencies/diseases/novel-coronavirus-2019
3. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis [Internet]. 2020 Feb 19 [cited 2020 Apr 24];0(0). Available from: https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(20)30120-1/abstract
4. An integrated national scale SARS-CoV-2 genomic surveillance network. Lancet Microbe [Internet]. 2020 Jun 2 [cited 2020 Jun 9];0(0). Available from:

https://www.thelancet.com/journals/lanmic/article/PIIS2666-5247(20)30054-9/abstract

5. CDC. SPHERES [Internet]. Centers for Disease Control and Prevention. 2020 [cited 2020 Jun 9]. Available from: https://www.cdc.gov/coronavirus/2019-ncov/covid-data/spheres.html
6. The COVID-19 Genomics UK (COG-UK) consortium and the Canadian COVID Genomics Network (CanCOGeN) launch new partnership [Internet]. [cited 2020 Jun 9]. Available from: https://www.genomecanada.ca/en/news/covid-19-genomics-uk-cog-uk-consortium-and-canadian-covid-genomics-network-cancogen-launch-new
7. Laboratory Guidelines for the Detection and Diagnosis of COVID-19 Virus Infection - PAHO/WHO | Pan American Health Organization [Internet]. [cited 2020 Jul 8]. Available from: https://www.paho.org/en/documents/laboratory-guidelines-detection-and-diagnosis-covid-19-virus-infection
8. Candido D et al, Evolution and epidemic spread of SARS-CoV-2 in Brazil. medRxiv 2020.06.11.20128249; doi: https://doi.org/10.1101/2020.06.11.20128249
9. 2019 Novel Coronavirus Resource (2019nCoVR), China National Center for Bioinformation [Internet]. 2019 Novel Coronavirus Resource (2019nCoVR). 2020 [cited 2020 Jun 30]. Available from: https://bigd.big.ac.cn/ncov/?lang=en
10. NGS-SA: Network for genomic surveillance in South Africa [Internet]. [cited 2020 Jul 24]. Available from: http://www.krisp.org.za/ngs-sa/ngs-sa_network_for_genomic_surveillance_south_africa/
11. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data – from vision to reality. Eurosurveillance. 2017 Mar 30;22(13):30494.
12. Cochrane G, Karsch-Mizrachi I, Takagi T, Sequence Database Collaboration IN. The International Nucleotide Sequence Database Collaboration. Nucleic Acids Res. 2016 Jan 4;44(D1):D48–50.
13. Allard MW, Strain E, Melka D, Bunning K, Musser SM, Brown EW, et al. Practical Value of Food Pathogen Traceability through Building a Whole-Genome Sequencing Network and Database. Kraft CS, editor. J Clin Microbiol. 2016 Aug;54(8):1975–83.
14. Kubota KA, Wolfgang WJ, Baker DJ, Boxrud D, Turner L, Trees E, et al. PulseNet and the Changing Paradigm of Laboratory-Based Surveillance for Foodborne Diseases. Public Health Rep Wash DC 1974. 2019 Dec;134(2_suppl):22S-28S.
15. Cook JA, Arai S, Armién B, Bates J, Bonilla CAC, Cortez MB de S, et al. Integrating Biodiversity Infrastructure into Pathogen Discovery and Mitigation of Emerging Infectious Diseases. BioScience [Internet]. 2020 [cited 2020 Jun 25]; Available from: https://academic.oup.com/bioscience/article/doi/10.1093/biosci/biaa064/5857068
16. World Health Organization (WHO) list of in-house-developed molecular assays for SARS-CoV-2 detection [Internet]. Available from: https://www.who.int/docs/default-source/coronaviruse/whoinhouseassays.pdf
17. WHO. Draft landscape of COVID-19 candidate vaccines [Internet]. 2020 [cited 2020 Jul 1]. Available from: https://www.who.int/publications/m/item/draft-landscape-of-covid-19-candidate-vaccines
18. In Silico evaluation of Diagnostic Assays [Internet]. Available from: https://covid19.edgebioinformatics.org/#/assayValidation
19. Díez-Fuertes F, Iglesias-Caballero M, Monzón S, Jiménez P, Varona S, Cuesta I, Zaballos A, Thomson MM, Jiménez M, Pérez JC, Pozo F, Pérez-Olmeda M, Alcamí J, Casas I. Phylodynamics of SARS-CoV-2 transmission in Spain. bioRxiv 2020.04.20.050039; doi:

https://doi.org/10.1101/2020.04.20.050039

20. Utkueri Y, Fer E, Bozlak E, Kutnu M, Kara NS, Yilmaz F. Comparison of SARS-CoV-2 variants with INSaFLU and galaxyproject [Internet]. BioHacker Xiv; 2020 May [cited 2020 Jun 9]. Available from: https://osf.io/9d3cz
21. Dudas G, Carvalho LM, Bedford T, Tatem AJ, Baele G, Faria NR, et al. Virus genomes reveal factors that spread and sustained the Ebola epidemic. Nature. 2017 Apr;544(7650):309–15.
22. Lai C-C, Wang J-H, Ko W-C, Yen M-Y, Lu M-C, Lee C-M, et al. COVID-19 in long-term care facilities: An upcoming threat that cannot be ignored. J Microbiol Immunol Infect. 2020 Jun;53(3):444–6.
23. Dyal JW. COVID-19 Among Workers in Meat and Poultry Processing Facilities — 19 States, April 2020. MMWR Morb Mortal Wkly Rep [Internet]. 2020 [cited 2020 Jul 1];69. Available from: https://www.cdc.gov/mmwr/volumes/69/wr/mm6918e3.htm
24. B.C. dentist dies after attending dental conference with COVID-19 outbreak [Internet]. Global News. [cited 2020 Jul 1]. Available from: https://globalnews.ca/news/6722164/dentist-dies-coronavirus-conference/
25. A family gathering led to 15 coronavirus cases. Experts say it’s ‘inevitable’ [Internet]. Global News. [cited 2020 Jul 1]. Available from: https://globalnews.ca/news/7053889/coronavirus-gatherings-cases-outbreaks/
26. Hodcroft EB. Preliminary case report on the SARS-CoV-2 cluster in the UK, France, and Spain. Swiss Med Wkly. 2020 24;150(9–10).
27. Caly L, Druce J, Roberts J, Bond K, Tran T, Kostecki R, et al. Isolation and rapid sharing of the 2019 novel coronavirus (SARS-CoV-2) from the first patient diagnosed with COVID-19 in Australia. Med J Aust. 2020;212(10):459–62.
28. Meredith LW, Hamilton WL, Warne B, Houldcroft CJ, Hosmillo M, Jahun AS, et al. Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated COVID-19: a prospective genomic surveillance study. Lancet Infect Dis [Internet]. 2020 Jul 14 [cited 2020 Jul 15];0(0). Available from: https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(20)30562-4/abstract
29. De Maio N, Walker C, Borges R, Weilguny L, Slodkowicz G, Goldman N. Issues with SARS-CoV-2 sequencing data [Internet]. Virological. 2020 [cited 2020 Jun 9]. Available from: https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473
30. Rayko M, Komissarov A. Quality control of low-frequency variants in SARS-CoV-2 genomes. bioRxiv. 2020 May 7;2020.04.26.062422.
31. Poon LLM, Leung CSW, Chan KH, Yuen KY, Guan Y, Peiris JSM. Recurrent mutations associated with isolation and passage of SARS coronavirus in cells from non-human primates. J Med Virol. 2005;76(4):435–40.
32. Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L, et al. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nat Biotechnol. 2011 May;29(5):415–20.
33. Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P, et al. The minimum information about a genome sequence (MIGS) specification. Nat Biotechnol. 2008 May;26(5):541–7.
34. Dugan VG, Emrich SJ, Giraldo-Calderón GI, Harb OS, Newman RM, Pickett BE, et al. Standardized Metadata for Human Pathogen/Vector Genomic Sequences. PLOS ONE. 2014 Jun 17;9(6):e99979.
35. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol. 2007 Nov;25(11):1251–5.

36. Schriml LM, Chuvochina M, Davies N, Eloe-Fadrosh EA, Finn RD, Hugenholtz P, et al. COVID-19 pandemic reveals the peril of ignoring metadata standards. Sci Data. 2020 Jun 19;7(1):188.
37. AusTrakka [Internet]. [cited 2020 Jul 1]. Available from: https://austrakka.net.au
38. GET Africa: Global Emerging Pathogens Treatment Consortium [Internet]. [cited 2020 Jul 24]. Available from: https://www.getafrica.org/
39. Baobab LIMS: An open source LIMS for biobanking developed by African and European Researchers. [Internet]. [cited 2020 Jul 24]. Available from: https://baobablims.org/
40. South African National Bioinformatics Institute (SANBI) [Internet] [cited 2020 Jul 1]. Available from: https://www.sanbi.ac.za/
41. Barrett T, Clark K, Gevorgyan R, Gorelenkov V, Gribov E, Karsch-Mizrachi I, et al. BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. Nucleic Acids Res. 2012 Jan;40(Database issue):D57–63.
42. NCBI. NCBI Pathogen Detection [Internet]. 2015 [cited 2020 Jun 17]. Available from: https://www.ncbi.nlm.nih.gov/projects/pathogens/
43. Home - Compare Europe [Internet]. https://www.compare-europe.eu. [cited 2020 Jun 9]. Available from: https://www.compare-europe.eu/