



Apr 09, 2021

PHA4GE contextual metadata SOP

In 1 collection

Emma Griffiths¹, Ruth E Timme², Nabil-Fareed Alikhan³, Duncan MacCannell⁴¹Simon Fraser University; ²US Food and Drug Administration; ³Quadram Institute Bioscience;⁴Centers for Disease Control and Prevention

1 Works for me dx.doi.org/10.17504/protocols.io.btpznmp6

GenomeTrakr StaPH-B 1 more workspace

Ruth Timme
US Food and Drug Administration

ABSTRACT

Contextual data curation is a critical part of public health pathogen genomic surveillance, as well as the data stewardship and data management that ensures the longevity and utility of the data. This protocol provides instructions and guidance for structuring information within the PHA4GE SARS-CoV-2 contextual data collection template in order to better enable harmonization of across datasets and systems. The template can be found [here](#).

The template file contains the **collection template**; a **field-level reference guide** containing examples, definitions and guidance; and a **controlled vocabulary** sheet which populates the pick lists in the template.

This protocol provides step-by-step instructions for populating the template, and also addresses a number of ethical, privacy and practical considerations that should be discussed with your data steward prior to any type of data sharing.

The appendices provide additional instructions and examples of how to curate sample type descriptions, and how to identify additional standardized terms should you need them.

Note: Data providers should review their jurisdictional and organization-specific data sharing policies prior to sharing any information contained in the template.

An **instructional video** providing an overview and demonstration of these materials, is also available.

The types of information that can be captured in the template include:

1. identifiers and repository accession numbers,
2. sample collection and processing,
3. host information,
4. host exposure information,
5. host reinfection information,
6. host vaccination information,
7. sequencing,
8. bioinformatics and QC metrics,
9. lineage and variant information,
10. pathogen diagnostic testing,
11. contributor acknowledgements.

A pdf version of this SOP is available here. [@ PHA4GE Contextual Data SOP 2.0.pdf](#)

For more information and/or assistance, contact datastructures@pha4ge.org.

EXTERNAL LINK

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

THIS PROTOCOL ACCOMPANIES THE FOLLOWING PUBLICATION

Griffiths EJ, Timme RE, Page AJ, Alikhan N, Fornika D, Maguire F, Mendes CI, Tausch SH, Black A, Connor TR, Tyson GH, Aanensen DM, Alcock B, Campos J, Christoffels A, Gonçalves da Silva A, Hodcroft E, Hsiao WW, Katz LS, Nicholls SM, Oluniyi PE, Olawoye IB, Raphenya AR, Vasconcelos ATR, Witney AA, MacCannell DR. The PHA4GE SARS-CoV-2 Contextual Data Specification for Open Genomic Epidemiology. Preprints; 2020 Aug. 9 [Epub ahead of print]. doi: 10.20944/preprints202008.0220.v1

DOI

dx.doi.org/10.17504/protocols.io.btpznmp6

EXTERNAL LINK

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

PROTOCOL CITATION

Emma Griffiths, Ruth E Timme, Nabil-Fareed Alikhan, Duncan MacCannell 2021. PHA4GE contextual metadata SOP. **protocols.io**
<https://dx.doi.org/10.17504/protocols.io.btpznmp6>

MANUSCRIPT CITATION please remember to cite the following publication along with this protocol

Griffiths EJ, Timme RE, Page AJ, Alikhan N, Fornika D, Maguire F, Mendes CI, Tausch SH, Black A, Connor TR, Tyson GH, Aanensen DM, Alcock B, Campos J, Christoffels A, Gonçalves da Silva A, Hodcroft E, Hsiao WW, Katz LS, Nicholls SM, Oluniyi PE, Olawoye IB, Raphenya AR, Vasconcelos ATR, Witney AA, MacCannell DR. The PHA4GE SARS-CoV-2 Contextual Data Specification for Open Genomic Epidemiology. Preprints; 2020 Aug. 9 [Epub ahead of print]. doi: 10.20944/preprints202008.0220.v1


COLLECTIONS ⓘ

 **SARS-CoV-2 NCBI submission workflow + guidance for structuring and releasing metadata**

KEYWORDS

SARS-CoV-2, metadata, contextual data, curation, data standard, genomic surveillance

LICENSE

 This is an open access protocol distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

CREATED

Mar 26, 2021

LAST MODIFIED

Apr 09, 2021

PROTOCOL INTEGER ID

48601

PARENT PROTOCOLS

Part of collection

[SARS-CoV-2 NCBI submission workflow + guidance for structuring and releasing metadata](#)

- 1 Download the file containing the collection template and reference guide from the following link:
<https://github.com/pha4ge/SARS-CoV-2-Contextual-Data-Specification>
- 2 Before you begin to curate your contextual data:
 1. Review your dataset
 2. Review the fields and values in the template.
 3. Review the field definitions and guidance in the template Reference Guide
- 3 Confirm the planned mapping of your data fields to those in the PHA4GE collection template with the data steward (e.g. your supervisor).

Note: Confirm the level of granularity of information that can be shared publicly and/or privately, with the data steward and/or your privacy officer. The most detailed information allowable should be included here. Different versions (detailed information vs general information) can be stored.

- 4 Populate the collection template with the information from your dataset.
 - Fields colour-coded yellow are considered mandatory. Fill these in first.
 - Fields colour-coded purple are strongly recommended. If you have permission, fill these fields in next.
 - Fields colour-coded white are optional, but still important. If you have permission, fill in these fields.
 - Use picklists where provided.
 - Ensure the data is stored safely with appropriate encryption

Note: Sometimes there will be constraints on what information can be shared, other times a field may not be applicable to your sample. Use the null values (controlled vocabulary indicating the reason why information is not provided) in the picklist to report missing data.

Table 1: "Required" fields for surveillance

Subsection	Required Fields
Sample Collection and Processing <i>Note: Consult your supervisor and/or data steward to evaluate whether the specimen collector sample ID is considered identifiable according to your institutional policies. If not considered identifiable, copy the sample ID into the "specimen collector sample ID" field in the collection template. If considered identifiable, provide the alternative sample ID. Be sure to keep a copy of the key in a safe location.</i>	specimen collector sample ID sample collected by sequence submitted by sample collection date geo_loc (country) geo_loc (province/territory) organism isolate
Host Information	host (scientific name) host disease
Bioinformatics and QC Metrics	consensus sequence software name consensus sequence software version

- 5 Use the SARS-CoV-2 contextual data Reference Guide to access field definitions, field-level guidance and examples.

See **Appendix A** for ethical and privacy considerations of contextual data.

See **Appendix B** for examples of how to structure sample descriptions.

If a desired term is not present in a picklist, you can search for a standardized term using the procedure in **Appendix C**.

- 6 Optional: Submit sequence data and corresponding contextual data to GISAID and/or an INSDC repository. See submission protocols and advice on preparing submission forms for more information.

Appendix A: Ethical, Practical, and Privacy Considerations

- 7 An effective and equitable response to the COVID-19 pandemic requires rapid and sustained international collaboration and data sharing. Many of the contextual data elements described in the PHA4GE SARS-CoV-2 contextual data specification are critical for effective public health surveillance and response. However, many of these same elements have ethical, practical, and privacy issues which must be considered before data can be shared. Data governance policies may vary between data types and jurisdictions, thus users of the specification should consult data stewards and privacy officers regarding organization-specific and jurisdiction-specific policies. Below, we highlight a series of common issues and provide suggestions for ways forward. The PHA4GE Reference Guide should be consulted for field-level guidance.

Note: This guidance is based on the experience of members of the PHA4GE working groups, and is not intended to apply to all situations and use cases. Decisions regarding implementation of the specification must ultimately be made

by the user in consultation with data providers and data stewards. If the intended use of the information collected is for research purposes, there will likely be many additional administrative and ethical requirements (e.g. Research Ethics Board (REB) review).

Identifiers and Repository Accession Numbers

Sharing consensus sequence and raw data, as well as contextual data, with public repositories enables tracking of global spread of the SARS-CoV-2 virus, phylodynamic analyses, development and improvement of diagnostics, and much more. Laboratories world-wide are sharing SARS-CoV-2 sequence and minimal contextual data with public repositories such as GISAID and the INSDC. When you share information with a public database, you will receive an accession number (a unique identifier in a database enabling the tracking of multiple versions of the data). If you have shared data with a public database, make sure to capture the accession numbers. GISAID will provide you with a single accession number. Make sure to record it. INSDC members (NCBI, ENA, DDBJ) may provide you with different accession numbers depending on what you share, and how. You can share assemblies and consensus sequences with GenBank (and its equivalents), raw data with Sequence Read Archive (SRA), and contextual data as a BioSample (see reference guide for further information). Information may be organized in BioProjects, and at a higher organizational level, Umbrella BioProjects. Make sure to record all of the applicable accession numbers.

Samples, libraries, patients, sequences (raw, processed, consensus etc) and so on can have many identifiers, especially if there is a division of labour or sharing of information across agencies and organizations. The specification has provided fields to capture many of those that are common, but may not capture all of the IDs you require. **It is essential to track IDs of original materials and information** to establish chain-of-custody and for follow-up, if necessary. It is better to track too many IDs than too few. If you require more fields to capture the IDs you need, add them. Some IDs are considered public health identifiable information (PHII). Make sure to check with the appropriate authorities whether the IDs you plan to share are considered identifiable information. If considered identifiable, you may need to create an alternative set of IDs. If you do, make sure to store the key in a safe and secure place.

Geographical Information

Geographical information (country, province/state/region, city, postal code, latitude/longitude etc) is very informative for tracking spread of the virus at different scales. Detailed geographical information for human clinical samples is often considered PHII depending on the number of cases in that locality, or may be specially regulated, and so must be abstracted before it can be shared. If the specification is being used for a sequencing project and detailed geographic information can be recorded, additional standardized fields such as geo_loc name (city), geo_loc name (county), host contact information (postal code) can be added to your collection template as needed. It is important to note that most geographic location fields in the specification **describe the sample**. Other fields have been provided to capture geo_loc information about the origin of the host and the likely country of exposure. Curators should ensure that the information they are entering correctly refers to the sample or the host. Before sharing data, especially with public repositories, it is important to ensure the data being submitted complies with the permitted level of granularity. Discuss this with the data steward. If sharing latitude and longitude coordinates, do not use the centre of the city/region/province/state/country or the location of your agency as a proxy, as this implicates a real location and is misleading.

The "host residence geo_loc (country)", "location of exposure geo_loc name (country)", and "host ethnicity" can be highly sensitive. If the information is shared and patients re-identified, it can have extreme consequences for the patient, the data collector, the data provider, and political relations. However, this information is important for characterizing risk, understanding transmission, and how the disease impacts some groups more than others (i.e. due to systemic health care inequity, poverty, racism etc). There may also be issues of equitable access and benefit sharing that should be considered for genomics data, particularly regarding Indigenous communities. Institutional, national and international resources regarding these issues should be consulted for best practices.

Date Information

Geographical and temporal information are key elements of infectious disease surveillance programs. Temporal information consists of dates e.g. sample collection date, sample received date, sample sequenced date, symptom onset date etc. Dates can be considered PHII on their own for human clinical samples, or in combination with other types of contextual data (e.g. geographical information), or in context of how many cases have been reported in a locality. Sharing "Sample collection date" along with sequence data is highly desirable, however If this date is considered identifiable, it is acceptable to add "jitter" to the collection date by adding or subtracting calendar days as required. Do not change the collection date in your original records. Furthermore, elements such as "sample collection date" are usually held by the institution that collected the original specimen (e.g. performed the diagnostic test). As such, you may require permission to acquire this information, or it may be difficult to attain due to other burdens on the data provider (workload, system access, manual curation requirements). Alternatively, "received date" may be used as

a substitute in the data you share.

Purpose of Sampling/Purpose of Sequencing

Sampling strategy can create biases in the data. A sample may be collected for one purpose, but sequenced for another (e.g. collected for diagnostic testing, but sequenced for surveillance of circulating lineages and variants). Information about why samples were collected and why they were selected for sequencing (i.e. random vs targeted sampling) can help inform epidemiological modelling and analyses. Standardized tags are available in the “purpose of sampling” field (e.g. Diagnostic Testing, Research, Surveillance) and in the “purpose of sequencing” field (e.g. Baseline surveillance (non-random sampling), Screening for Variants of Concern (VoC), Cluster/Outbreak Investigation). Free text fields are also available for providing extra information about sampling and the selection of samples for sequencing called “purpose of sampling details” and “purpose of sequencing details”. A number of standardized phrases are also suggested in the Reference Guide for describing different common surveillance priorities.

Host Information

Outside of specifying the species’ scientific or common name, human host information is almost always considered PHII. Patient information is usually collected at the time of specimen collection (e.g. diagnostic test) using a case report form, and held by the institution that collected the original specimen. You will more than likely require permission to acquire this information, or it may be difficult to attain due to other burdens on the data provider (workload, system access, manual curation requirements).

“Host age” and “Host gender” are regularly collected for most surveillance programs and can be used to characterize case definitions, and for linkage between lab and epidemiological data. On their own, this information may not be considered PHII, however, they may be considered identifiable information when combined with other contextual data such as collection date and geographical location. Abstracting age information by using age binning is acceptable in the specification. Suggested age bins are as follows: 0-9 years, 10-19 years, 20-29 years, 30-39 years, 40-49 years, 50-59 years, 60-69 years, 70-79 years, 80-89 years, 90-99 years, and 100+ years.

Information about whether an individual was asymptomatic or symptomatic, and known health outcomes can be recorded in the “host health state” field, while information about whether they were hospitalized, medically isolated, self-quarantining etc can be recorded under “host health state details”. Sharing of these types of information in aggregated form is often permissible, however sharing of these data types at an individual level, along with information about signs and symptoms, pre-existing conditions and risk factors, and complications, are usually restricted.

Host Exposure Information

The context of pathogen exposure is very important for understanding transmission chains and for determining public health actions. An individual may be exposed through direct or indirect contact with an infected person at an event, in a particular location (exposure setting) through a variety of contexts (host roles), or through travel. This information is usually highly sensitive.

Methods Information

Methodological information, such as sampling and experimental design, laboratory procedures, bioinformatic processing, and quality control metrics, are crucial information to understand the context and limitations of analyses. Capturing as much well-structured information regarding your methods, and storing it in a centralized place (or single document) helps to future-proof the data as well as the work that went into collecting, processing, analyzing and interpreting the data. Capturing methodological information also enables better reproducibility, and increases quality control. The specification provides many fields for capturing experimental design, protocols, and scientific metrics. It is strongly recommended that as much of that information be captured and stored as possible.

Null Values

The International Nucleotide Database Collaboration (INSDC) have created standardized [missing/null value reporting language](#) to be used where a value of an expected format for sample metadata reporting can not be provided. This controlled vocabulary has been adopted in this specification, and takes into account different types of constraints (i.e. Not Applicable, Missing, Not Collected, Not Provided, Restricted Access). Users are strongly encouraged to always provide as much information as possible in the collection template, however, if missing/null value reporting is required, users are asked to use a term with the finest granularity for their situation.

Note: NCBI accepts all null values. ENA will accept any other null value besides “Missing”.

V. Appendix B: Describing your sample.

8 Why, how and when samples are collected can impact analyses of sequence data. In determining how a virus spreads,

it is critical to track temporal and geographical information. It is also important to capture as much data provenance (who contributed it, where it came from, how it was generated) as possible. Different sampled materials or sampling processes may contain higher viral loads or produce better results, and differences in sampling protocols and practices should be accounted for (e.g. to understand sampling effects on interpreting a genomics-based cluster, to identify mutations due to viral passage in the lab). A number of recommended and optional fields are provided to capture sampling methods ("purpose of sampling", "specimen processing", "lab host", "passage number", "passage method"). We highly recommend including information regarding whether the virus was passaged, and how. Seven fields have been 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. throat swab should be recorded:

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

e.g. combined nasopharynx/oropharynx samples should be recorded:

specimen processing	host (scientific name)	host (common name)	host disease	anatomical part	collection device
Specimens Pooled	Homo sapiens	Human	COVID-19	Nasopharynx; Oropharynx	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

e.g. tissue from a bat (*Platyrrhinus lineatus*) in a cave should be recorded:

Host (common name)	Host (scientific name)	host disease	anatomical part	environmental site
Bat	Platyrrhinus lineatus	Not Applicable	Tissue	Cave

e.g. particulates from air filter should be recorded:

environmental material	collection method
Particulate Matter	Air Filtration

VI. Appendix C: How to Find Standardized Terms

- 9 Pick lists of standardized vocabulary will be made available in the collection template, and will be refined based on user feedback. If a desired term cannot be found in a pick list, the instructions below outline steps to identify additional standardized terms.

Identifying Standardized Terms

1. Go to the [EBI Ontology look-up service](#). Links to appropriate ontologies within the service are available in the SOP and template reference guide.
2. Enter your term of interest in the search bar. The closest matching results will be displayed.
3. Select the term that is the best match and copy and paste it into your collection template in the appropriate column.
4. If you have difficulty finding a term that matches your input, consider entering synonyms of your desired term. If you can't find a term in the ontology suggested in the SOP, try expanding your search by entering your term in the general search bar at <https://www.ebi.ac.uk/ols>.

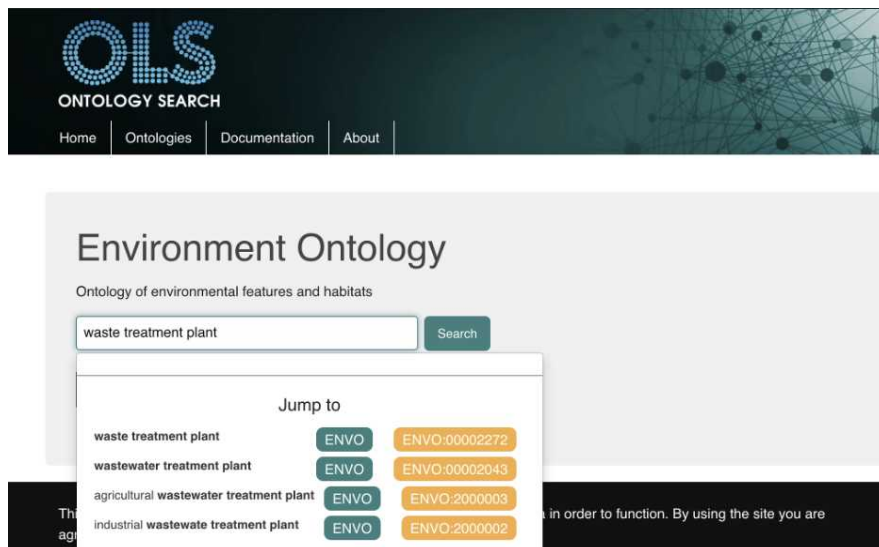
Example:

Term: *waste treatment plant*

This contextual data describes an environmental location.

The "environmental location" guidance tells us to use the EnvO ontology to source standardized terms.

So we go to <https://www.ebi.ac.uk/ols/ontologies/envo>, and enter "waste treatment plant" in the search bar.



Many search results are returned, but we can see a term “waste treatment plant” that matches our term.

Open the Vocabulary tab in the collection template.

Copy the term and paste it in alphabetical order into the “environmental site” list in your collection template.

Then click “Data” in the tool menu at the top of the page, followed by “Data validation”. Under “Allow”, it should say “List”.

Highlight all of the vocabulary (not including the column header) in the “environmental site” list, then click “OK”. This step will expand the source of the terms that provide the pick list.

Return to the Template tab and search the “environmental site” pick list and you should find your newly added term.

A tutorial about how to search for new standardized terms and how to add them to the template is also available.

For more information and/or assistance, contact datastructures@pha4ge.org.