Alberta Microbiota Repository (AMBR)
AMBR_SOP_1.0 Contextual Data Curation

1. Background

Genomic and genetic characterization of microorganisms in samples from different environments (e.g. natural or built environments, hosts, laboratory culture) are powerful tools for taxonomic identification and elucidating genotype-to-phenotype relationships. How sequence data is generated (sample context, growth and isolation conditions, sequencing techniques, bioinformatic processes) can have profound impacts on the data itself, its use, interpretations of the data, and any subsequent biological insights that result from analyses. Contextual data is the sample, laboratory, clinical, epidemiological, and methods information that enables the interpretation of sequence data. Contextual data is often captured using free text in databases and spreadsheets, and so often contains a high degree of variability in data structure (fields/terms/formats) and the meaning and organization of vocabulary, and can contain jargon, errors, and differences in granularity. By structuring contextual data using community standards such as minimum information checklists and ontologies, this information can be more easily understood and used by both humans and computers, and can be more easily reproduced, compared and reused for different types of analyses.

The Alberta Microbiota Repository (AMBR), led by the Harrison Lab at the University of Calgary, is an interdisciplinary study aimed at using 16S sequencing as part of a culturomics platform to identify antibiotic potentiators from the natural products of microbiota. The AMBR team has partnered with the Centre for Infectious Disease Genomics and One Health (CIDGOH) at Simon Fraser University to standardize the contextual data in its isolate repository. CIDGOH specializes in the development of ontologies and data standards for pathogen genomics in public health and food safety. CIDGOH's data specifications and harmonization tools have been used during the COVID-19 pandemic for Canadian and international SARS-CoV-2 data sharing, as well as other initiatives and laboratory networks such as the FDA's GenomeTrakr for foodborne pathogen surveillance. CIDGOH implements many of its data specifications via a data curation, validation and automated transformation tool called the DataHarmonizer. The DataHarmonizer provides ontology-based fields and terms to users in a spreadsheet-style text editor, and enables browser-based data curation. Ontologies are collections of controlled vocabulary that are arranged in a hierarchy, where all the terms are linked using logical relationships. Ontologies are open source, community developed, and meant to represent "universal truth" as much as possible (so not tied to one organization's vocabulary use case). Ontologies encode synonyms, which enables mapping between the specific languages used by different organizations, and every term in the ontology is assigned a globally unique and persistent identifier. Using ontology terms to standardize AMBR contextual data not only helps make data more interoperable by using a common language, it also helps to make contextual data FAIR (Findable, Accessible, Interoperable, Reusable).

To better harmonize AMBR contextual data, the DataHarmonizer now provides an AMBR-specific template containing standardized fields, pick lists of controlled vocabulary and prescribed formats. The standardized fields and terms have been sourced from a variety of ontologies (e..g GenEpiO, NCBITaxon, EnvO, Uberon, OBI etc). The template is accompanied

by different supporting materials such as Field and Term reference guides (which provide definitions and additional specific guidance) and this curation SOP.

2. Specification Design Principles

The AMBR DataHarmonizer specification is intended to capture identifiers linked to samples, isolates, and sequences in order to establish chains of custody, improve “institutional memory” with good record-keeping, and to better enable follow-up in case more information is necessary. The specification also aims to capture sample descriptors, pertinent host information, isolate characteristics and growth conditions, sequencing and bioinformatics analysis methods, and contributor details.

The specification is divided into seven parts that organize fields into the following sections: “Database identifiers”, “Sample collection and processing”, “Strain and isolate information”, “Host information”, “Sequencing”, “Bioinformatics and QC metrics”, and “Contributor acknowledgement”. As the AMBR collection focuses on isolates rather than samples, the central identifier in the “isolate ID”. “Specimen collector sample ID”s may be linked to isolates, and while this is good practice for traceability, it is not required. When (dates), where (geographical and contextual locations), how (collection methods) and why (criteria for collection) a sample is collected is important information for interpreting data and for understanding potential bias. Providing the most granular descriptions of samples is preferred, including what was sampled (i.e. the material - be it anatomical part, body product or environmental material), where the sampled material was taken from (anatomical site of a host or an environmental site) and how it was collected (via a specific device, vessel or technique such as necropsy). How specimens and isolates are processed can affect downstream sequencing, and so the specification provides fields for capturing specimen processing as well as isolation methods.

Hosts are considered living things that harbour microbes - which here includes animals (such as humans and horses) and plants. Hosts can be referred to by their scientific names (e.g. Equus cabellus) and by their common names (e.g. horse), and the specification provides fields to distinguish between the different types of nomenclatures.

As methodology can also impact sequencing results, the specification provides several fields for tracking sequencing instruments, protocols and critical reagents such as primers used to generate 16s rRNA amplicons. The specification also tracks bioinformatics processes and tools such as reference database names and version numbers, analysis metrics (percent coverage and identity), and top hit search results.

To help ensure that contributions by different researchers and lab personnel are acknowledged, names of key individuals involved in the research can be included in the isolate contextual data records.

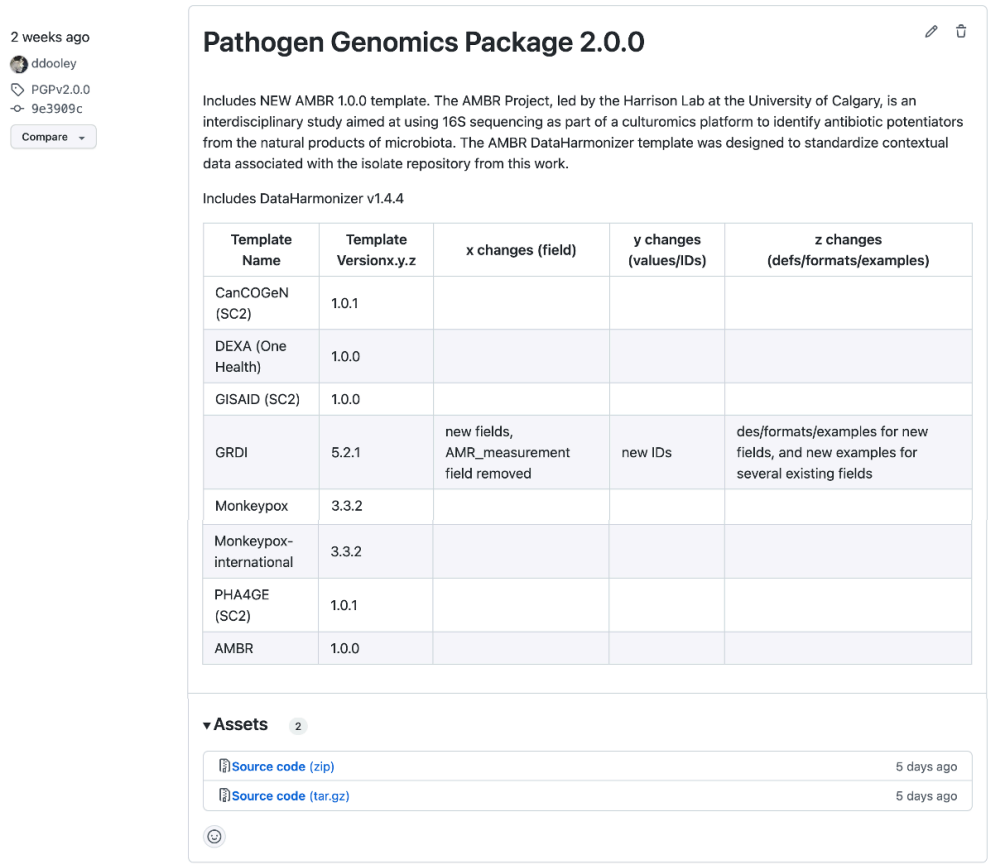
While ontologies can contain extensive vocabulary, it is not always useful to have long picklists of values for different fields. As such, the specification contains vocabulary customized to the scope of the samples being captured. However, the specification is intended to evolve and grow over time with changing data needs, and so vocabulary can be added as need by making term requests by emailing curators of the Genomic Epidemiology Ontology (GenEpiO)

via email (info@genepio.org) or via the GenEpiO GitHub issuetracker (https://github.com/GenEpiO/genepio/issues).

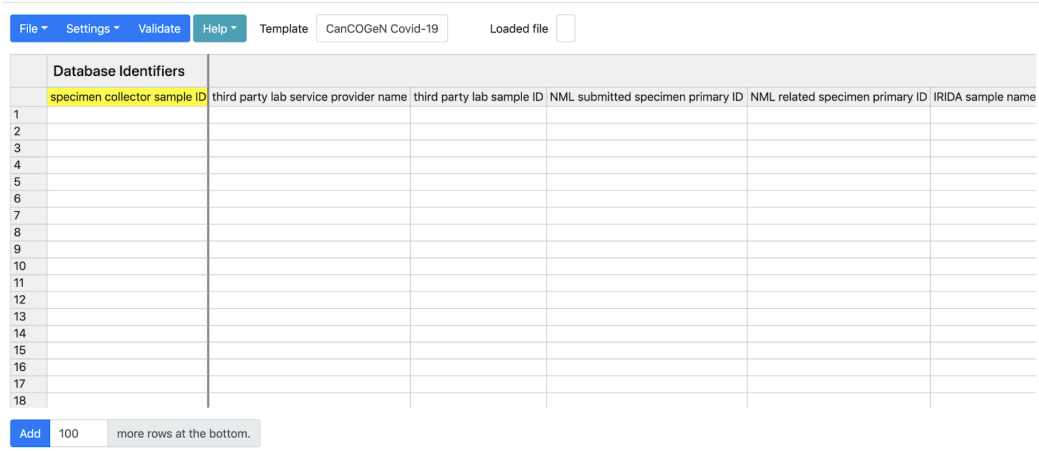
3. How to Curate AMBR Contextual Data Using the DataHarmonizer

The following procedure and annexes outline the steps for using the DataHarmonizer to curate and validate AMBR contextual data.

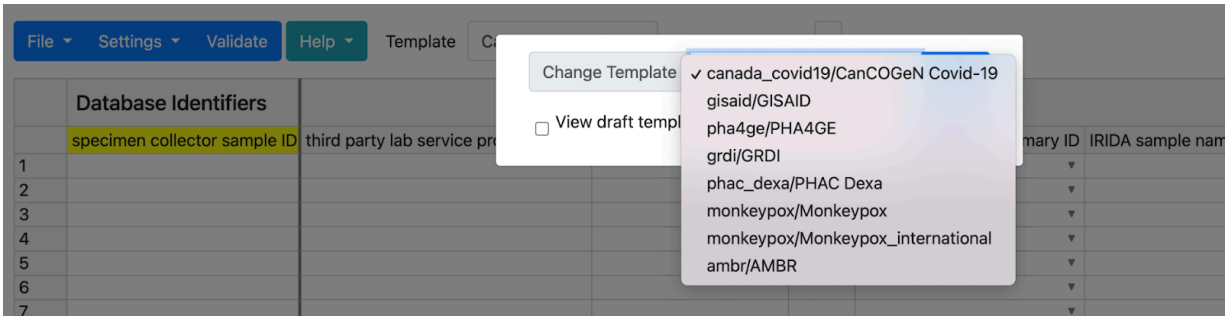
	Action
1	Download the zip file ("Source code (zip))" containing The DataHarmonizer application from the following link: https://github.com/cidgoh/pathogen-genomics-package/releases



Extract the zip file's contents, and navigate into the extracted folder. Open **main.html**. The validator application will open in your default browser. It should look like this:



The DataHarmonizer enables contextual data harmonization for different pathogens and projects. Select the AMBR template by selecting “ambr/AMBR” from the **Template** menu beside the **Help** button.

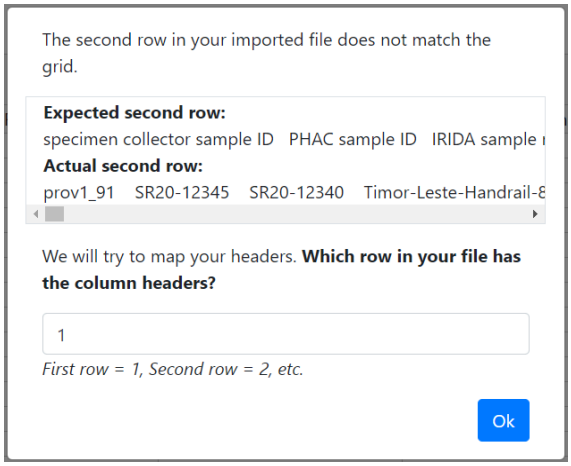


Data can be entered into the validator application manually, by typing values into the application’s spreadsheet, or data can be imported from local `xlsx`, `xls`, `tsv` and `csv` files.

To import local data, click **File** on the top-left toolbar, and then click **Open**. To enter data in a new file, click **File** on the top-left toolbar, and then click **New**. Data entered into the spreadsheet can be copied and pasted.

Note: Only files containing the headers expected by the DataHarmonizer can be opened in the application.

If you are missing the first row, you will get the following warning:

	 *Resolve by declaring "1" as the row in which your column headers reside.*
2	Before you begin to curate sample metadata: ● Review your data ● Review the fields in the template of the Validator application ● Review the field descriptions in the SOP Appendix
3	Familiarize yourself with DataHarmonizer functionality by reviewing the "**Getting Started**". To access "Getting Started", click on the green **Help** button on the top-left toolbar, then click **Getting Started**. Definitions, examples and further guidance are available by double clicking on the field headers, or by using the "**Reference Guide**". To access the "Reference Guide" click on the **Help** button, then click **Reference Guide**.
4	Confirm mapping of your data fields to those in the harmonized template with the data steward (e.g. your supervisor). *Note: Examples of how data should be entered can be found in the Master AMBR file.*
5	Enter data into the validator spreadsheet. ● Hide non-required fields (colour-coded purple and white/grey) by clicking **Settings** on the top-left toolbar, followed by clicking on **Show Required Columns** (colour-coded in yellow). ● Double click in the field headers to see definitions and detailed guidance as needed (or consult Appendix A). ● Jump to a specific field header by clicking **Settings** on the top-left toolbar, followed by clicking on **Jump to**, then select the field header of the column you would like to view from the drop down list. ● Populate the validator template with the information from your dataset. ● Use picklists when provided. ● A value must be entered for every *required field* in each row. If data is missing or not collected, **choose a null value from the picklist**.

    - Not Applicable
    - Missing
    - Not Collected
    - Not Provided
    - Restricted Access
- Free text can be provided when picklists are not available.
- For filling an entire column with the same data, use the **Fill Column** function. Click **Settings**, followed by **Fill Column**. Type in the name of the desired field, followed by the value that should be used to fill every row in that column. Then click **OK**.

If a desired term is not present in a picklist, contact Emma Griffiths at ega12@sfu.ca.

Note: Sometimes a field may not be applicable to your isolate. Use the null values (controlled vocabulary indicating the reason why information is not provided) in the picklist to report missing data.

Required fields are organized into subsections (see **Appendix A** for required field definitions and guidance):

Subsection	Required Fields
Database Identifiers	isolate ID
Sample Collection and Processing *Note: Seven fields have been introduced to capture different kinds of anatomical and environmental samples, as well as collection methods. Populate only the fields that pertain to your sample - provide null values for the fields that are not applicable. Select the appropriate value from the pick list provided (consult the reference guide for further support). Provide the most granular information available.*	organism anatomical material anatomical part body product environmental material environmental site collection device collection method
Strain and Isolate Information	strain taxonomic identification method taxonomic identification method details incubation temperature value incubation temperature unit

<table>
<tr><td></td><td>
<table>
<tr><td></td><td>isolation medium
isolate storage location</td></tr>
<tr><td>Host Information</td><td>host (common name)</td></tr>
<tr><td>Sequencing</td><td>sequencing instrument
amplicon pcr primer list</td></tr>
<tr><td>Bioinformatics and QC Metrics</td><td>reference accession
reference database name
reference database version
coverage (percentage)
sequence identity percentage
top-hit taxon determination
top-hit strain determination
trimmed ribosomal gene sequence</td></tr>
</table>
</td></tr>
<tr><td>6</td><td>Validate the entered data by clicking on the Validate button on the top-left toolbar.

Missing information and invalid entries in required fields will be highlighted in red.
● Observe invalid rows by clicking Settings in the top-left toolbar, and then clicking on Show invalid rows.
● Address errors systematically by clicking the Next Error button. When all errors have been corrected, the Next Error button will disappear.
● Observe valid rows by clicking Settings in the top-left toolbar, and then clicking on Show valid rows.
● Return view to all rows by clicking Settings in the top-left toolbar, and then clicking on Show all rows.

Note: Row viewing options only appear after a validation attempt has been made.</td></tr>
<tr><td>7</td><td>Address any invalid data that was flagged in red in the template.
● Pale Red = Incorrect data format
● Dark Red = Required data missing

Note: It is possible to export incomplete or invalid data. Make sure to review any errors prior to exporting.</td></tr>
<tr><td>8</td><td>Save a version-controlled copy of the AMBR Master file and store it somewhere safe.</td></tr>
<tr><td>9</td><td>Additional Information:</td></tr>
</table>

	A local copy of the **Standard Operating Procedure (SOP)** is included in every download of the DataHarmonizer. To access it, click on the green **Help** button on the top-left toolbar, then click **SOP**. The latest version of the SOP is published online and accessible via a web browser at all times.

Appendix A: Required Field Definitions and Guidance

Field definitions for required fields, as well as guidance and examples, are provided below. This information has been sourced from the DataHarmonizer reference guide. Guidance for strongly recommended and optional fields can be found in the reference guide. For access to information on non-required fields, refer to “Procedure - Action 3”**.**

<u>Database Identifiers</u>

isolate ID

The user-defined identifier for the isolate, as provided by the laboratory that originally isolated the isolate.

Provide the identifier created by the lab for the organism after isolation. This value maps to the "Strain ID#" in the Alberta Microbiota Repository (AMBR) Master file.

e.g. SA01

<u>Sample Collection and Processing</u>

organism

Taxonomic name of the organism.

Provide the confirmed taxonomic name of the species. This value maps to the "Recommended identification" in the Alberta Microbiota Repository (AMBR) Master file.

e.g. Staphylococcus aureus

<u>Describing the material and/or site sampled.</u>

anatomical material

A substance obtained from an anatomical part of an organism e.g. tissue, blood.

Provide a descriptor if an anatomical material was sampled. Use the picklist provided in the template. If a desired term is missing from the picklist, contact emma_griffiths@sfu.ca. If not applicable, do not leave blank. Choose a null value. Information for populating this field may be available in the "Source of Isolation" field in the Alberta Microbiota Repository (AMBR) Master file.

e.g. Wound tissue (injury)

anatomical part

An anatomical part/location of an organism e.g. oropharynx.

Provide a descriptor if an anatomical part was sampled. Use the picklist provided in the template. If a desired term is missing from the picklist, contact emma_griffiths@sfu.ca. If not applicable, do not leave blank. Choose a null value. Information for populating this field may be available in the "Source of Isolation" field in the Alberta Microbiota Repository (AMBR) Master file.

e.g. Nasal cavity

body product

A substance excreted/secreted from an organism e.g. feces, urine, sweat.

Provide a descriptor if a body product was sampled. Use the picklist provided in the template. If a desired term is missing from the picklist, contact emma_griffiths@sfu.ca. If not applicable, do

not leave blank. Choose a null value. Information for populating this field may be available in the "Source of Isolation" field in the Alberta Microbiota Repository (AMBR) Master file.
e.g. Feces

environmental material
A substance or object obtained from the natural or man-made environment e.g. soil, water, sewage.
Provide a descriptor if an environmental material was sampled. Use the picklist provided in the template. If a desired term is missing from the picklist, contact emma_griffiths@sfu.ca. If not applicable, do not leave blank. Choose a null value. Information for populating this field may be available in the "Source of Isolation" field in the Alberta Microbiota Repository (AMBR) Master file.
e.g. Bandage

environmental site
An environmental location may describe a site in the natural or built environment e.g. contact surface, metal can, hospital, wet market, bat cave.
Provide a descriptor if an environmental site was sampled. Use the picklist provided in the template. If a desired term is missing from the picklist, contact emma_griffiths@sfu.ca. If not applicable, do not leave blank. Choose a null value. Information for populating this field may be available in the "Source of Isolation" field in the Alberta Microbiota Repository (AMBR) Master file.
e.g. Hospital

collection device
The instrument or container used to collect the sample e.g. swab.
Provide a descriptor if a device was used for sampling. Use the picklist provided in the template. If a desired term is missing from the picklist, contact emma_griffiths@sfu.ca. If not applicable, do not leave blank. Choose a null value. Information for populating this field may be available in the "Source of Isolation" field in the Alberta Microbiota Repository (AMBR) Master file.
e.g. Swab

collection method
The process used to collect the sample e.g. phlebotomy, necropsy.
Provide a descriptor if a collection method was used for sampling. Use the picklist provided in the template. If a desired term is missing from the picklist, contact emma_griffiths@sfu.ca. If not applicable, do not leave blank. Choose a null value. Information for populating this field may be available in the "Source of Isolation" field in the Alberta Microbiota Repository (AMBR) Master file.
e.g. Biopsy

<u>Strain and Isolate Information</u>
strain
The strain identifier.
Provide the strain of the isolate. This value maps to the "Strain" in the Alberta Microbiota Repository (AMBR) Master file.
e.g. CL10

taxonomic identification method
The type of planned process by which an organismal entity is associated with a taxon or taxa.
Provide the type of method used to determine the taxonomic identity of the organism by selecting a value from the pick list. For the AMBR Project, the "16S ribosomal gene sequencing assay" value will be the most appropriate. If the information is unknown or cannot be provided, leave blank or provide a null value.
e.g. 16S ribosomal gene sequencing assay

taxonomic identification method details
The details of the process used to determine the taxonomic identification of an organism.
Provide the criteria used for 16S sequencing taxonomic determination by selecting a value from the pick list. These criteria are specific to the AMBR project and so do not correspond with standardized criteria in any ontology. The pick list is strictly for providing consistency in records rather than implementing community data standards. If another method was used for the taxonomic determination, leave blank. This value maps to the information stored in the "ID Category*" field in the Alberta Microbiota Repository (AMBR) Master file.
e.g. Species-level ID: >99.3% identity and unambiguous match to one type (T) sequence in a curated database

incubation temperature value
An environmental datum specifying the temperature at which an organism or organisms were incubated for the purposes of growth on or in a particular medium.
Provide the temperature at which the isolate was isolated. This value maps to the information stored in the "Incubation temperature" field in the Alberta Microbiota Repository (AMBR) Master file.
e.g. 37

isolation medium
An isolation medium is a culture medium which has the disposition to encourage growth of particular bacteria to the exclusion of others in the same growth environment.
Select the temperature unit from the pick list. This value maps to the information stored in the "Incubation media" field in the Alberta Microbiota Repository (AMBR) Master file.
e.g. Brain heart infusion (BHI)

incubation temperature unit
An environmental datum specifying the temperature unit at which an organism or organisms were incubated for the purposes of growth on or in a particular medium.
Select the isolation medium from the pick list. This value maps to the information stored in the "Incubation temperature" field in the Alberta Microbiota Repository (AMBR) Master file.
e.g. Degree Celsius

isolate storage location
An isolate datum specifying the location of where an isolate is stored e.g. in a particular freezer, on a particular shelf.
Enter the freezer storage location of the isolate as the "freezer number-shelf number-box number-unit number" e.g. FR1-R3-B1-S01. This value maps to the information stored in the "Spot code" in the Alberta Microbiota Repository (AMBR) Master file.

e.g. FR1-R3-B1-S01

cellular respiration type
An isolate datum specifying the type of cellular respiration process used by the organism.
Select the respiration type from the pick list. This value maps to the information stored in the "Aerobic/Anaerobic" field in the Alberta Microbiota Repository (AMBR) Master file.
e.g. Aerobic respiration

Host Information
host (common name)
The commonly used name of the host.
Common name is required if there was a host. Both common name and scientific name can be provided, if known. Use terms from the pick lists in the template. Hosts can be animals (including humans) and plants. Examples of common names are “Human” and “Canola plant”. Examples of scientific names are “Homo sapiens” and “Equus caballus”. If the sample was environmental, select "Not Applicable". Information for populating this field may be available in the "Source of Isolation" field in the Alberta Microbiota Repository (AMBR) Master file.
e.g. Human

Sequencing
sequencing instrument
The model of the sequencing instrument used.
Select a sequencing instrument from the picklist provided in the template.
e.g. MinIon

amplicon pcr primer list
An information content entity specifying a list of primers used for amplicon sequencing.
Select the primers used to generate the ribosomal 16S or 23S amplicon for sequencing from the pick list. This value maps to the information in the "Primers Used for sequencing" field Alberta Microbiota Repository (AMBR) Master file.
e.g. 27F;1492R

Bioinformatics and QC Metrics
reference accession
An identifier that specifies an individual sequence record in a public sequence repository.
Enter the EZBioCloud gene accession that most closely matches the sequence being analyzed. This value maps to the information in the "Accession No(s)." field in the Alberta Microbiota Repository (AMBR) Master file.
e.g. FR821777

reference database name
An identifier of a biological or bioinformatics database.
Select the reference database name from the pick list. For the AMBR Project, the reference database will be EZBioCloud.
e.g. EZBioCloud

reference database version

The version of the database containing the reference sequences used for analysis.
Enter the sequence search date as the version of EZBioCloud used. Record the date in ISO 8601 format i.e. YYYY_MM_DD. This value maps to the information in the "Search date" field in the Alberta Microbiota Repository (AMBR) Master file.
e.g. 2021-05-23

coverage (percentage)
The percentage of the reference sequence covered by the sequence of interest.
Enter the completeness value. Do not include any symbols e.g. %. This value maps to "Completeness (%)" in the Alberta Microbiota Repository (AMBR) Master file.
e.g. 98.2

sequence identity percentage
Sequence identity is the number (%) of matches (identical characters) in positions from an alignment of two molecular sequences.
Enter the identity value. Do not include any symbols e.g. %. This value maps to "Similarity (%)" in the Alberta Microbiota Repository (AMBR) Master file.
e.g. 99

top-hit taxon determination
The taxon derived from the top hit in search results produced from a sequence similarity comparison.
Enter the EZBioCloud taxon best-hit. This value maps to the information in the "Top-hit taxon (taxa)" field in the Alberta Microbiota Repository (AMBR) Master file.
e.g. Staphylococcus argenteus

top-hit strain determination
The strain designation derived from the top hit in search results produced from a sequence similarity comparison.
Enter the EZBioCloud strain best-hit. This value maps to the information in the "Top-hit strain(s)" field in the Alberta Microbiota Repository (AMBR) Master file.
e.g. MSHR1132(T)

trimmed ribosomal gene sequence
The results of a data transformation of sequence data in which (e.g., low quality) read bases are removed to produce a trimmed ribosomal RNA sequence.
Enter the sequence of the trimmed ribosomal gene sequence. This value maps to the sequence in the "Trimmed Ribosomal Sequence" field in the Alberta Microbiota Repository (AMBR) Master file.
e.g.TGCAAGTCGAGCGAACGGACGAGAAGCTTGCTTCTCTGATGTTAGCGGCGGACGSGTG
AGTAACACGTGGATAACCTACCTATAAGACTGGGATAACTTCGGGAAACCGGAGCTAATACC
GGATAATATTTTGAACCGCATGGTTCAAAAGTGAAAGACGGTCTTGCTGTCACTTATAGATGG
ATCCGCGCTGCATTAGCTAGTTGGTAAGGTAACGGCTTACCAAGGCAACGATGCATAGCCG
ACCTGAGAGGGTGATCGGCCACACTGGAACTGAGACACGGTCCAGACTCCTACGGGAGG
CAGCAGTAGGGAATCTTCCGCAATGGGCGAAAGCCTGACGGAGCAACGCCGCGTGAGTGA
TGAAGGTCTTCGGATCGTAAAACTCTGTTATTAGGGAAGAACATATGTGTAAGTAACTGTGCA
CATCTTGACGGTACCTAATCAGAAAGCCACGGCTAACTACGTGCCAGCAGCCGCGGTAATA

CGTAGGTGGCAAGCGTTATCCGGAATTATTGGGCGTAAAGCGCGCGTAGGCGGTTTTTTAA
GTCTGATGTGAAAGCCCACGGCTCAACCGTGGAGGGTCATTGGAAACTGGAAAACTTGAG
TGCAGAAGAGGAAAGTGGAATTCCATGTGTAGCGGTGAAATGCGCAGAGATATGGAGGAAC
ACCAGTGGCGAAGGCGACTTTCTGGTCTGTAACTGACGCTGATGTGCGAAAGCGTGGGGA
TCAAACAGGATTAGATACCCTGGTAGTCCACGCCGTAAACGATGAGTGCTAAGTGTTAGGGG
GTTTCCGCCCCTTAGTGCTGCAGCTAACGCATTAAGCACTCCGCCTGGGGAGTACGACCGC
AAGGTTGAAACTCAAAGGAATTGACGGGGACCCGCACAAGCGGTGGAGCATGTGGTTTAAT
TCGAAGCAACGCGAAGAACCTTACCAAATCTTGACATCCTTTGACAACTCTAGAGATAGAGC
CTTCCCCTTCGGGGGACAAAGTGACAGGTGGTGCATGGTTGTCGTCAGCTCGTGTCGTGA
GATGTTGGGTTAAGTCCCGCAACGAGCGCAACCCTTAAGCTTAGTTGCCATCATTAAGTTGG
GCACTCTAAGTTGACTGCCGGTGACAAACCGGAGGAAGGTGGGGATGACGTCAAATCATC
ATGCCCCTTATGATTTGGGCTACACACGTGCTACAATGGACAATACAAAGGGCAGCGAAACC
GCGAGGTCAAGCAAATCCCATAAAGTTGTTCTCAGTTCGGATTGTAGTCTGCAACTCGACTA
CATGAAGCTGGAATCGCTAGTAATCGTAGATCAGCATGCTACGGTGAATACGTTCCCGGGTC
TTGTACACACCGCCCGTCACACCACGAGAGTTTGTAACACCCGAAGCCGGTGGAGTAACCT
TTTAGGAGCTAGCCGTCGAAG

Appendix B: Structuring Sample Descriptions (Examples)

Several examples are provided below which illustrate how to structure common sample descriptions.

e.g. lake water from Lake Louise, AB should be recorded:

original sample description	geo_loc_name (state/province/territory)	geo_loc (site)	environmental site	environmental material
lake water from Lake Louise, AB	Alberta	Lake Louise	Lake	Water

e.g. nasal swab from an American cystic fibrosis patient as part of the CF123-01 collection project should be recorded:

original sample description	sample collection project name	geo_loc_name (country)	host (scientific name)	host (common name)	host disease	anatomical part	collection device
nasal swab from an American cystic fibrosis patient as part of the CF123-01 collection project	CF123-01	United States of America	Homo sapiens	Human	Cystic fibrosis	Nasal cavity	Swab

e.g. leaf from a willow tree should be recorded:

original sample description	host (common name)	anatomical part
leaf from a willow tree	Willow tree	Leaf

e.g. gauze associated with a wound should be recorded:

original sample description	host (scientific name)	host (common name)	environmental material	anatomical material

gauze associated with a wound	Homo sapiens	Human	Gauze	Wounded tissue (injury)

Appendix C: Null Value Definitions

Not Applicable
Information is inappropriate to report, can indicate that the standard itself fails to model or represent the information appropriately.

Missing
Information was known to be recorded in the past, but the observed value cannot be located or retrieved for some reason.

Not Collected
Information of an expected format was not given because it has not been collected.

Not Provided
Information of an expected format was not given, a value may be given at the later stage.

Restricted Access
Information exists but can not be released openly because of privacy concerns.

Source:
International Nucleotide Sequence Database Collaboration (INSDC) Missing Value Reporting Terms (2017-2018).
ENA Training Modules: https://ena-docs.readthedocs.io/en/latest/submit/samples/missing-values.html

Appendix D: Field Mapping from AMBR Master Copy 2022-08-10 to the AMBR DataHarmonizer Template

Below, are one-to-one field mappings identifying the relationships between the fields in the previous AMBR Master inventory to the AMBR DataHarmonizer Specification. Please note that there are additional fields in the DataHarmonizer specification that do not exist in the previous AMBR Master inventory. The additional fields are present to improve the capture of sample and bioinformatics analysis provenance.

AMBR Master	AMBR DataHarmonizer Specification

Strain ID#	**isolate ID**
Recommended identification	**organism**
Label ID	**alternative isolate ID**
Source of isolation	**original sample description** **sample collection project name** **purpose of sampling** **geo_loc_name (country)** **geo_loc_name (state/province/territory)** **geo_loc_name (city)** **geo_loc_name (site)** **anatomical material** **anatomical part** **body product** **environmental material** **environmental site** **collection device** **collection method** **host (common name)** **host (scientific name)** **host disease**
Isolation media	**isolation medium**
Incubation temperature	**incubation temperature value** **incubation temperature unit**
Aerobic/Anaerobic	**cellular respiration type**
Strain Spot Code	**isolate storage location**
ID Category*	**taxonomic identification method details**
Top-hit taxon (taxa)	**top-hit taxon determination**
Top-hit strain(s)	**top-hit strain determination**
Similarity (%)	**sequence identity percentage**
Completeness (%)	**coverage (percentage)**
Variation ratio	**sequence identity (variance ratio)**
Accession No(s).	**reference accession**
Method	**No equivalent field**

Primers Used for sequencing	**amplicon pcr primer list**
Comment	**bioinformatics analysis details**
Search date	**reference database version**
Sequence Batch #	**library ID**
Trimmed Ribosomal Sequence	**trimmed ribosomal gene sequence**

Revision History

Version	Date	Writer	Description of Change
1.0	Jan 26 2023	Emma Griffiths	Created protocol