

Oct 30, 2024

Version 4

## GenFS Metadata Cleanup Challenge protocol V.4

DOI

<https://dx.doi.org/10.17504/protocols.io.rm7vzj6prlx1/v4>



Ruth Timme<sup>1</sup>, Martin Shumway<sup>2</sup>, Candace Bias<sup>3</sup>, Maria Balkey<sup>3</sup>, Tina Pfefer<sup>4</sup>

<sup>1</sup>US Food and Drug Administration; <sup>2</sup>National Center for Biotechnology Information; <sup>3</sup>US FDA-HFP; <sup>4</sup>US FDA

GenomeTrakr

Tech. support email: [genomeTrakr@fda.hhs.gov](mailto:genomeTrakr@fda.hhs.gov)



Ruth Timme

US Food and Drug Administration

### Create & collaborate more with a free account

Edit and publish protocols, collaborate in communities, share insights through comments, and track progress with run records.

Create free account

OPEN  ACCESS



**DOI:** <https://dx.doi.org/10.17504/protocols.io.rm7vzj6prlx1/v4>

**Protocol Citation:** Ruth Timme, Martin Shumway, Candace Bias, Maria Balkey, Tina Pfefer 2024. GenFS Metadata Cleanup Challenge protocol. protocols.io <https://dx.doi.org/10.17504/protocols.io.rm7vzj6prlx1/v4> Version created by **Candace Hope Bias**

**License:** This is an open access protocol distributed under the terms of the **Creative Commons Attribution License**, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

**Protocol status:** Working

**We use this protocol and it's working**

**Created:** October 29, 2024

**Last Modified:** October 30, 2024

**Protocol Integer ID:** 111188

**Keywords:** ncbi for the genfs metadata cleanup challenge exercise, genfs metadata cleanup challenge protocol this protocol, genfs metadata cleanup challenge protocol, genfs metadata cleanup challenge exercise, metadata cleanup challenge, submitting bulk biosample, bulk biosample, ncbi, fda, finalized template, final step, update, metadata

## Abstract

This protocol provides guidance for submitting bulk BioSample updates to NCBI for the GenFS Metadata Cleanup Challenge exercise.

Table of Contents:

- Timeline/Workflow for the 2024 Metadata Cleanup Challenge
- Preparing the update file
- Making corrections and updates within the file
- Submitting the finalized template

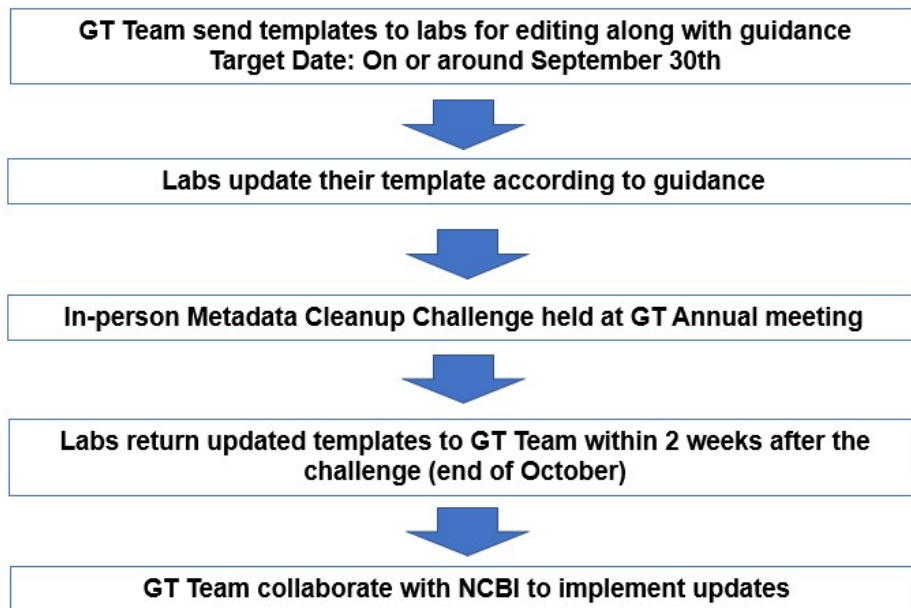
## Version History:

**V4:** addition of a training video

**V3:** minor edits made to correct typos. Addition of final step to email FDA and NCBI when you've completed all the steps.

## Timeline/Workflow for the 2024 Metadata Cleanup Challenge

1



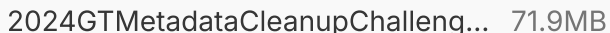
## 2 Overview and scope of exercise

**Timeframe:** focus on submissions between September 2023-August 2024. Your lab is welcome to include submission outside this timeframe.

**Scope:** review and curate entries for the following attributes, which are core requirements within the One Health Enteric package:

### Focus on these attributes:

- source\_type
- collected\_by
- sequenced\_by
- project\_name
- host (for human and animal isolates)
- food\_origin (for food isolates)
- isolation\_source



71.9MB

## Guidance for preparing the update file

**3 Steps 4 and 5 describe the required format for the bulk update file.**

#### 4 For labs in the GenomeTrakr network:

We have generated a bulk metadata update template for every laboratory that has a BioProject linked to the GenomeTrakr umbrella BioProject at NCBI (**PRJNA593772**).

Pick up your metadata template here: [DOCS: metadata hackathon/2024 Hackathon Template Pickup.](#)

**Instructions for generating your own template (if your lab's BioProject is not linked to the GenomeTrakr Umbrella):**

**Navigate to SRA Run Selector:** <https://www.ncbi.nlm.nih.gov/Traces/study/>

Enter your BioProject accession(s), then click the "Metadata" button in the download box.

**SRA Run Selector**

Accession PRJNA692474 Q Search

**Filters List**

- ☐ attribute\_package
- ☐ AvgSpotLen
- ☐ Bases
- ☐ BioProject
- ☐ BioSampleModel
- ☐ Bytes
- ☐ Center Name
- ☐ collected\_by
- ☐ Genus
- ☐ geo\_loc\_name\_country
- ☐ geo\_loc\_name\_country\_co
- ☐ Host
- ☐ Host\_disease
- ☐ IFSAC\_Category
- ☐ IFSAC+\_Category
- ☐ Instrument

**Common Fields**

|                    |                                         |
|--------------------|-----------------------------------------|
| Consent            | PUBLIC                                  |
| Assay Type         | WGS                                     |
| DATASTORE filetype | FASTQ, RUN.ZQ, SRA                      |
| DATASTORE provider | GS, NCBI, S3                            |
| DATASTORE region   | gs.us-east-1, ncbi.public, s3.us-east-1 |
| LibraryLayout      | PAIRED                                  |
| LibrarySource      | GENOMIC                                 |
| Platform           | ILLUMINA                                |
| version            | 1                                       |

**Select**

|                 | Runs | Bytes    | Bases     | Download                                                |
|-----------------|------|----------|-----------|---------------------------------------------------------|
| <b>Total</b>    | 241  | 66.33 Gb | 136.36 Gb | <span>Metadata</span> or <span>Accession List</span>    |
| <b>Selected</b> | 0    | 0        | 0         | <span>Metadata</span> or <span>Accession List</span> or |

The download metadata files will be a comma-delimited txt file with both the BioSample + SRA metadata. Open this in Excel. Remove all the metadata columns not relevant for this exercise, move the BioSample accessions to the first column, then proceed. See the example template for guidance on which columns to keep:



Example Template.xlsx

If you see that any of the isolates listed in your template should be part of another organization's submissions (i.e. PulseNet), please notify the GenomeTrakr team by emailing [genometrakr@fda.hhs.gov](mailto:genometrakr@fda.hhs.gov).

## 5 Prepare the update file:

- Updates should be put into one bulk update file per lab, or per lab/organism, or some other aggregation.
- **Rows:** The first row, or header, must contain the column names. Each subsequent row must contain exactly one BioSample and its updatable fields, or attributes. Each row must contain exactly the same number of columns.
- **Columns:** The column names should be attribute names (**full name or harmonized name**) included in the One Health Enteric package. The first column should contain the BioSample accession (eg SAMN123456789)
- **File format:** tab-delimited text file using the file suffix .tsv or .txt. If you use Microsoft Excel for editing, export the final template into this format. Take care that dates are correctly exported from the spreadsheet.

### FYI:

- Only those columns where at least one record has a replacement value need to be supplied.
- If a field is blank in the update file, and has an actual value in NCBI BioSample, that will be taken as "replace current value with NULL".
- If a field in a record already has a value in NCBI BioSample that will not be changed, then the field should be filled in with its current value.
- If a field is new to the BioSample, it will be added. A blank field whose attribute does not currently exist in the BioSample record will not be added.

### 5.1 What fields/attributes can be included in the bulk update file?

- collected\_by
- geo\_loc\_name
- food\_origin
- host
- project\_name
- sequenced\_by
- source\_type

- isolation\_source
- collection\_date
- purpose\_of\_sampling
- env\_local\_scale
- env\_medium
- animal\_env
- env\_broad\_scale
- facility\_type
- intended\_consumer
- food\_type\_processed
- food\_processing\_method

## 5.2 What fields **CANNOT** be updated in the bulk update file?

DO NOT include changes to the following attributes in this bulk update file. Changes to the following fields can be requested separately by writing to **pd-help@ncbi.nlm.nih.gov** directly. Changes to the following attributes affect multiple resources at NCBI and updates to them need to be coordinated.

- **biosample\_acc/BioSample** – This is the primary key of the entire BioSample system - Keep this accession in the first column, but DO NOT EDIT OR UPDATE ENTRIES.
- **bioproject\_acc/BioProject** - changes to linked bioProject, linked SRA, or linked Assembly.
- **strain** or **isolate**
- **sample\_name**
- **attribute package** – This requires validation of all the fields together.
- **center name** – This is a property of SRA and cannot be changed using the bulk update channel.
- **organism** - Change to the BioSample species (the identification of the isolate).
- **Salmonella serovar/serotype** - changes to organism species or sub-species names (ex Salmonella enterica ⇒ Salmonella enterica subsp. enterica serovar Infantis)
- Any kind of SRA to BioSample, BioSample to BioProject, or SRA to BioProject mapping.
- Any of the **computed fields** in Pathogen Detection (epi\_type, min\_same, min\_diff, computed\_types, or amr/virulence analysis outputs). These attributes cannot be updated by the record owner.

For updating any of these attributes, send a TSV file of proposed changes to **pd-help@ncbi.nlm.nih.gov** as prepared above.

## Guidance for making corrections and updates within the file



6 **Steps 7-9:** These steps cover required fields. In your Excel template, ensure there's an entry in these columns for every row. If information is missing, then choose one of the null terms, "Not Applicable, Not Collected, Not Provided, Missing, or Restricted Access".

7 **source\_type**

Review your entries, ensure every record has an source\_type entry. Make corrections where needed.

human  
animal  
food  
environmental  
other  
Not Applicable  
Not Collected  
Not Provided  
Missing  
Restricted Access

NCBI and US pathogen surveillance coordinators have a *strict* controlled vocabulary for this attribute, only the above terms are allowed.

8 **project\_name**

Review the entries in this column, make corrections where needed and populate all missing fields. For US surveillance, please choose the coordinating body that best the isolate.

GenomeTrakr  
GenomeTrakr; LFFM-FY1  
GenomeTrakr; LFFM-FY2  
GenomeTrakr; LFFM-FY3  
GenomeTrakr; LFFM-FY4  
GenomeTrakr; LFFM-FY5  
NARMS  
NARMS Cecal  
NARMS Retail Meat  
PulseNet  
USDA-FSIS  
Vet-LIRN  
NAHLN  
USDA-ARS



If you would like to create another term to communicate membership in another project or network, feel free to do that! Enter your new term directly into the update template. If you would like this term added to the picklist, send this term to **genomeTrakr@fda.hhs.gov**, and we'll add it to the next update.

Include more than one term? Separate with "; ", for example, "GenomeTrakr; USDA-ARS".

## 9 **collected\_by** and **sequenced\_by**

Review the entries in both of these columns, make corrections where needed, and populate any missing fields.

Ensure terms are standardized across all your records. Check the **sequenced\_by** picklist in the current One Health Enteric package file for your standardized laboratory name.

Send updates, corrections, or additions to your laboratory name to **genomeTrakr@fda.hhs.gov** and we'll make corresponding updates to the picklist terms.

- 10 **Steps 11-13:** For this exercise we're focused on two conditionally required attributes:
1. **host** should be populated for isolates derived from human or animal samples
  2. **food\_origin** should be populated for isolates derived from food products, or other commercial products sampled for pathogens (medical products, cosmetics, tattoo ink, etc).

Use the **source\_type** column to filter for these sample types, first for animal/human samples (step 11 + 12), then for food samples (step 13).

## 11 **host**

### **\*FOR HUMAN and ANIMAL ISOLATES ONLY\***

**Host** is a required for host associated human and animal isolates.

Use the **source\_type** column to identify human and animal isolates (filter in Excel for human and animal).

Use the entries in the **isolation\_source** column to help determine what the **host** entry should be.

Where possible, enter a scientific or binomial name, for example *Homo sapiens* or *Bos taurus*. If scientific name is unknown, use a common name recognized by the **NCBI taxonomy database** (porcine, bovine, etc).

When **host** is completely unknown, provide one of NCBI's null values:

Not Applicable  
Not Collected  
Not Provided  
Missing  
Restricted Access

## 12 isolation\_source

**\*\*For human and animal isolates: remove all taxonomic references included in `isolation_source`.\*\***  
This information should solely reside in **host**.

Example edit to both **isolation\_source** and **host**:

**Pathogen: environmental/food/other sample from *Salmonella enterica***

|             |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |                                                                                                                                                                                                                                                                                                                                                                                                     |
|-------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Identifiers | BioSample: SAMN07774651; SRA: SRS3721856; CFSAN: CFSAN070305                                                                                                                                                                                                                                                                                                                                                                                                                                      |                                                                                                                                                                                                                                                                                                                                                                                                     |
| Organism    | <a href="#">Salmonella enterica</a><br><small>cellular organisms; Bacteria; Pseudomonadota; Gammaproteobacteria; Enterobacterales; Enterobacteriaceae; Salmonella</small>                                                                                                                                                                                                                                                                                                                         |                                                                                                                                                                                                                                                                                                                                                                                                     |
| Package     | <a href="#">Pathogen: environmental/food/other; version 1.0</a>                                                                                                                                                                                                                                                                                                                                                                                                                                   |                                                                                                                                                                                                                                                                                                                                                                                                     |
| Attributes  | <b>Interagency Food Safety Analytics Collaboration (IFSAC) category</b><br><b>strain</b><br><b>source type</b><br><b>geographic location</b><br><b>collected by</b><br><b>collection date</b><br><b>isolate name alias</b><br><b>latitude and longitude</b><br><b>isolation source</b><br><b>project name</b><br><b>sequenced by</b><br><b>ontological term</b><br><b>attribute_package</b><br><b>PublicAccession</b><br><b>ProjectAccession</b><br><b>Species</b><br><b>Genus</b><br><b>Host</b> | veterinary clinical/research  cow<br>CFSAN070305<br>animal<br><a href="#">Nigeria</a><br>University of Ibadan<br>2012<br>DE 52<br>missing<br>stool ( <u>Bos taurus</u> )<br>GenomeTrakr<br>FDA Center for Food Safety and Applied Nutrition<br>bos taurus:NCBITAXON_9913,<br>feces:UBERON_0001988<br>environmental/food/other<br>CFSAN070305<br>PRJNA186035<br>enterica<br>Salmonella<br>Bos taurus |
| BioProject  | <a href="#">PRJNA186035</a> Salmonella<br>Retrieve <a href="#">all samples</a> from this project                                                                                                                                                                                                                                                                                                                                                                                                  |                                                                                                                                                                                                                                                                                                                                                                                                     |

## 13 food\_origin

**\*\*FOR FOOD OR OTHER PRODUCTS ONLY\*\***

Sort or filter on the **source\_type** column to identify the **"food"** isolates.

Food and other products have two attributes describing geographic location information:



**geo\_loc\_name:** geographic location where the sample was physically collected

**food\_origin:** geographic *origin* of food product or other product sampled for pathogens

Check location data in **geo\_loc\_name:**

- If the location information in **geo\_loc\_name** reflects the *state or country of origin* for the food product (e.g. "India", reflecting an imported food product sampled in the US), move this geographic information to **food\_origin**.
- If your isolates contain the location where the sample was collected (**e.g. port of entry, US state of grocery store, etc**), leave **geo\_loc\_name** as is. Populate **food\_origin** with the country or state of origin.

If **food\_origin** is unknown, provide one of NCBI's null values:

Not Applicable

Not Collected

Not Provided

Missing

Restricted Access

## OPTIONAL: bring your records up to OHE standards

- 14 Review and populate the other conditionally required fields for OHE sub-packages. Download **latest version of the One Health Enteric package** for picklist terms.

Filter on **source\_type** to locate isolates belonging to the four OHE sub-packages. The terms listed under the sub-packages are the conditionally mandatory attributes for those subpackages.

**animal samples**, source\_type = animal  
animal\_env

**food product samples**, source\_type = food  
intended\_consumer  
food\_processing\_method

**facility inspection samples**, source\_type = environmental  
facility\_type  
food\_type\_processed

**farm/environment samples**, source\_type = environmental  
env\_broad\_scale  
env\_medium  
env\_local\_scale



## Finalize the update file and submit the bulk template!

15 Check for duplications - ensure there are no double BioSample entries.

15.1 Bring your template to the October 16th, 2024 GenomeTrakr metadata cleanup challenge - finalize edits at the hackathon.

15.2 Remove columns that you're not updating. Ensure that the BioSample accession is retained in Column 1 and strain is kept for tracking the update.

### **Keep**

Biosample  
strain

### **Remove**

package name  
Title  
Center\_Name  
Release\_Date  
Bioproject  
Run  
sample\_name

15.3 Save as a tab-delimited .tsv file and transfer the .tsv file to protocols.io:  
**DOCS: metadata hackathon/2024 Hackathon Template Submission**

15.4 Send an email to **genomeTrakr@fda.hhs.gov** and **pd-help@ncbi.nlm.nih.gov**, alerting us that you've completed the update. Include your filename in the email. If NCBI has questions as they are processing the update, they will reach out to you directly.